Федеральное агентство по образованию
Государственное образовательное учреждение
высшего профессионального образования
ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Э. В. Ивантер
А. В. Коросов

ЭЛЕМЕНТАРНАЯ БИОМЕТРИЯ

Учебное пособие

Рекомендуется Учебно-методическим объединением по биологическому образованию вузов России в качестве учебного пособия для студентов биологических специальностей

2-е издание

Петрозаводск
Издательство ПетрГУ
2010

УДК 578.087.1
ББК 28.08:22.172
И 228

Рецензент:
В. А. Илюха, ведущий научный сотрудник ИБ КарНЦ РАН

*Печатается по решению
редакционно-издательского совета
Петрозаводского государственного университета*

Ивантер, Э. В., Коросов, А. В.
И 228 Элементарная биометрия : учеб. пособие / Э. В. Ивантер, А. В. Коросов. — Петрозаводск: Изд-во ПетрГУ, 2010. — 104 с.
ISBN 978-5-8021-1112-3

Книга служит элементарным пособием для практического применения вариационной статистики в биологических исследованиях.

В краткой, доступной форме на конкретных примерах рассмотрены приемы количественной обработки материалов биологических наблюдений и экспериментов. Приводятся алгоритмы статистических расчетов, показаны принципы биологической интерпретации математических показателей, раскрыты основы статистического оценивания, проверки гипотез, применения методов корреляционного, регрессионного, дисперсионного анализов.

Книга рассчитана на биологов различного профиля, студентов, аспирантов, научных и практических работников, преподавателей вузов и школ, специалистов сельского и лесного хозяйства, здравоохранения и ветеринарии.

УДК 578.087.1
ББК 28.08:22.172

ISBN 978-5-8021-1112-3

© Ивантер, Э. В., Коросов, А. В. 2010
© Петрозаводский государственный университет, 2010

ВЫБОРКА

Биометрическое исследование в центр внимания всегда ставит *выборку – множество значений случайной величины, совокупность вариант, набор чисел*; отдельная варианта – это объект, несущий качественный или числовой признак. Термин «выборка» указывает на процесс выбора части из чего-то большего, в данном случае – на процесс получения ограниченного количества значений из генеральной совокупности. *Генеральная совокупность – это множество всех вариант определенного типа* (выборка бесконечного размера). Чаще всего получить все возможные значения в принципе невозможно. Поэтому судить о генеральной совокупности приходится, исследуя выборки, – по части составлять представление о целом.

Признак

Варианта качественно или количественно выражает признак данного объекта исследования (полученного при данном уровне фактора внешней среды вполне определенным методом). Признак (свойство, показатель, величина, характеристика, переменная) – любая информация о наблюдаемом объекте, выраженная качественно или определенная количественно. В рамках вариационной статистики признаки выступают в роли случайной величины. Случайная величина – численная характеристика, принимающая те или иные заранее точно не известные значения. Несмотря на то, что точное описание поведения случайной величины получить нельзя, математическая статистика позволяет выполнить вероятностное описание.

Существует целый ряд методов регистрации признаков биологических объектов.

Качество (нечисловой дискретный признак) – простой, непосредственный, чувственный способ регистрации фактов; это статус, сезон, таксон, цвет, плотность, тип действия и пр. Значения таких признаков выражаются словами или символами, они не имеют количественного содержания и выражают принадлежность данного объекта к определенной обширной группе объектов (зеленый, январь, ♀, ♪).

Балл (оценка) – дискретный полуколичественный признак, численная характеристика объекта, присвоенная в соответствии с внешней заранее принятой шкалой баллов. Во время оценки объект соотносится с этими критериями и ему присваивается соответствующий балл. Баллы не обладают свойствами чисел, в частности, балл 4 не в два раза больше балла 2, для них арифметические операции применять нельзя. Для баллов многие выборочные параметры (средние, дисперсии и др.) не будут обладать свойствами статистических параметров, их нельзя статистически сравнивать, например, с помощью критерия Стьюдента. Корректно будет характеризовать выборки балльных оценок лишь с помощью частотных распределений, моды, размаха изменчивости. Для статистической обработки балльных оценок требуются *непараметрические* методы.

Количество (число) – дискретный (счетный) количественный признак (число натурального ряда), характеризующий множество однородных объектов, черт, деталей строения, состав (например, число эмбрионов у самки, число жаберных тычинок у рыб, число тычинок в цветке, число деревьев на пробной площадке). Отдельную варианту получают, подсчитав число неких дискретных черт строения у отдельного объекта или в пробе. *Проба – ограниченная совокупность разнокачественных объектов, которая характеризуется числом объектов одного определенного качества, это значение играет роль одной варианты выборки*. Получая серию проб, мы осуществляем перевод качественных признаков в количественные.

Промер (ряд дробных или рациональных чисел) – непрерывный (мерный) количественный признак, характеризующий свойства объектов с помощью различных относительных количественных шкал – температурной, весовой, размерной, объемной и т. п. Отдельная варианта получает количественную характеристику выраженности данного признака у данного объекта (в пределах точности метода): температуру тела, его размеры, уровень глюкозы в крови и т. д. Большинство методов статистики разработано для исследования именно таких непрерывных признаков (параметрические методы).

Варьирование

Основная особенность выборки как множества значений случайной величины – это отличие отдельных вариант друг от друга, явление *изменчивости*, варьирования, появления отличий между отдельными вариантами.

Биологу важно знать обычные причины варьирования. Один из источников, эндогенный, – это индивидуальные отличия по *статусу* и по *состоянию*. Например, животные одного возраста различны индивидуально, генетически, т. е. по статусу. Кроме того, каждое из них в разные годы, сезоны, время суток имеет разные морфофизиологические характеристики, т. е. отличается по состоянию. В наиболее точных науках (токсикология, биохимия, молекулярная биология) стремятся с помощью химической чистоты постановки опытов и выведения чистых линий подопытных животных убрать все мешающие причины «избыточного» варьирования.

Другой источник отличий между вариантами – факторы внешней среды, т. е. условия проведения наблюдений, среда существования объекта, возможная причина, определяющая текущее состояние объекта. Часто говорят про факторы эндогенные, внутренние (статус, способ существования объекта), и экзогенные, внешние (среда, условия существования объекта). Фактор всегда есть активное, действующее начало, признак – его результат, последствие. Факторы, влияющие на значения вариант, различаются по своей природе. Если фактор влияет на все варианты выборки постоянно и примерно одинаково, он называется систематическим (или доминирующим). Если фактор непостоянен, влияет на варианты не одинаково, с разной силой, он определяется как случайный. Эти рассуждения дают *модели варианты*:

$$x_i = x_{дом.} \pm x_{случ.},$$

где x_i – измеренное значение варианты,

i – индекс варианты ($i = 1, 2, ..., n$),

n – объем (общее количество вариант) выборки,

$x_{дом.}$ – суммарный вклад j доминирующих факторов,

$x_{случ.}$– суммарный вклад k случайных факторов.

С методической точки зрения при наблюдениях или в эксперименте самым важным оказывается обязательная *регистрация* максимально возможного числа факторов (как внешних, так и внутренних). Тогда появляется возможность исследовать их раздельное действие на объект, поскольку существуют методы, которые позволяют из многокомпонентной среды вычленять эффекты действия отдельных факторов (особенно работоспособны дисперсионный, регрессионный и компонентный анализы).

При самом широком варьировании признаков разброс значений выборки не бесконечно широк, он ограничен неким диапазоном и тяготеет к определенному общему значению. Эти свойства статистических совокупностей – варьирование, но в ограниченном диапазоне, – позволяют предложить для описания две группы величин: оценку центрального значения диапазона (среднюю, моду или медиану) и оценку размаха варьирования (лимит, дисперсию, стандартное отклонение). Определение этих значений выполняется после построения вариационного ряда.

Построение вариационного ряда

Любое статистическое исследование должно начинаться с установления характера распределения изучаемых признаков. *Распределение – это соотношение между значениями случайной величины и частотой их встречаемости.* Бóльшая повторяемость одних значений по сравнению с другими заставляет задумываться о причинах наблюдаемых процессов. Если значения признака откладывать по оси абсцисс, а частоты их встречаемости – по оси ординат, то можно построить *гистограмму*, *частотную диаграмму*, удобную для целей иллюстрации и исследования.

Основой для построения гистограммы служит *вариационный ряд – представленный в виде таблицы ряд значений изучаемого признака, расположенных в порядке возрастания с соответствующими им частотами их встречаемости в выборке.*

Начнем с примера изучения плодовитости серебристо-черных лисиц, которое дало следующие результаты (число щенков на самку): 5 5 6 5 5 6 4 4 4 5 6 4 6 6 4 6 4 5 5 8 5 3 6 5 5 5 5 5 6 3 6 4 6 4 6 2 5 6 5 3 7 6 3 4 6 8 6 3 5 5 6 5 4 3 8 4 7 5 4 3 1 6 5 3 4 5 6 7 4 4 6 5 6 4 6 5.

Плодовитость, x	Частота, a
1	1
2	1
3	8
4	16
5	23
6	21
7	3
8	3

Для дискретного признака (такова плодовитость) построение вариационного ряда обычно не представляет сложности, достаточно подсчитать встречаемость конкретных значений.

Гистограмма, построенная по данным о плодовитости лисиц (рис. 2), сразу же обнаруживает характерное поведение случайной величины – высокие частоты встречаемости значений в центре распределения и низкие – по периферии.

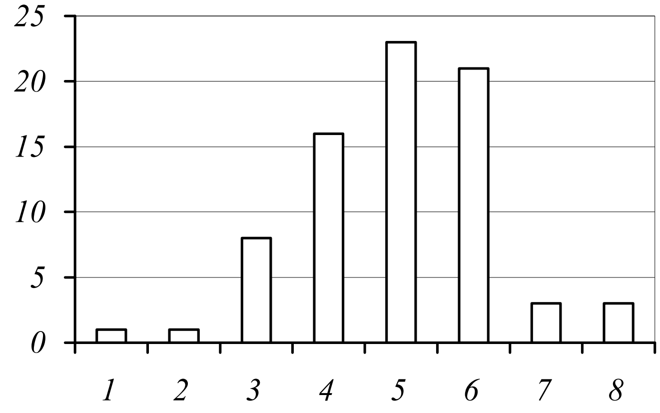


Рис. 2. *Распределение плодовитости лисиц*

Если же изучаемый признак непрерывен (таковы размерно-весовые характеристики), то для построения вариационного ряда сначала весь диапазон изменчивости признака разбивается на серию равных интервалов (классов вариант), затем подсчитывают, сколько вариант попало в каждый интервал. Число классов для больших выборок ($n > 100$) должно быть не менее 7 и не более 12, их оптимальное число можно приблизительно определить по эмпирической формуле:

$k = 1 + 3.32 \cdot \lg(n)$, где *n – объем выборки* (*число вариант в выборке*).

Составим для примера вариационный ряд для непрерывного признака – по данным о весе 63 взрослых землероек (г):

9.2	11.6	8.1	9.1	10.1	9.6	9.3	9.7	9.9	9.9	9.6
7.6	10.0	9.7	8.4	8.6	9.0	8.8	8.6	9.3	11.9	9.3
9.2	10.2	11.2	8.1	10.3	9.2	9.8	9.9	9.3	9.1	9.4
9.6	7.3	8.3	8.8	9.2	8.0	8.6	8.8	9.0	9.5	9.1
8.5	8.8	9.7	11.5	10.5	9.8	10.0	9.4	8.7	10.0	7.9
8.6	8.7	9.1	8.2	9.2	9.4	8.8	9.8			

1) Все операции могут быть выполнены вручную. Вначале следует определить объем выборки $n = 63$.

2) Рассчитать пределы *размаха изменчивости* значений, *лимит – разность между максимальным и минимальным значениями*:

$Lim = x_{max} - x_{min} = 11.9 - 7.3 = 4.6$.

3) Найти число классов вариационного ряда по формуле:

$k = 1 + 3.32 \cdot \lg(63) = 6.973811 \approx 7$.

4) Найти длину интервала dx (допустимо округление):

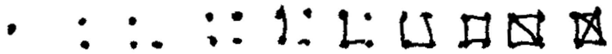
$dx = Lim / k = 4.6 / 7 \approx 0.7$.

5) Установить границы классов; в качестве первой границы имеет смысл взять округленное минимальное значение: $x_{min} = 7$.

6) Вычислить центральное значение признака в каждом классе; исходным берется значение центра первого интервала; для первого класса 7–7.7, для второго – 7.8–8.4…

7) Произвести разноску вариант в соответствующие классы с подсчетом их числа методом конверта (табл. 2):

1 2 3 4 5 6 7 8 9 10 .



Теперь данные можно представить графически, в виде полигона частот (ломаной кривой) или гистограммы (столбиками) (рис. 3).

Таблица 2

Классы	Центр классового интервала	Подсчет частот	Частоты, а
7–7.7	7.35	:	2
7.8–8.4	8.05		7
8.5–9.1	8.75		18
9.2–9.8	9.45		22
9.9–10.5	10.15		10
10.6–11.2	10.85	•	1
11.3–11.9	11.55	:.	3
Сумма			63

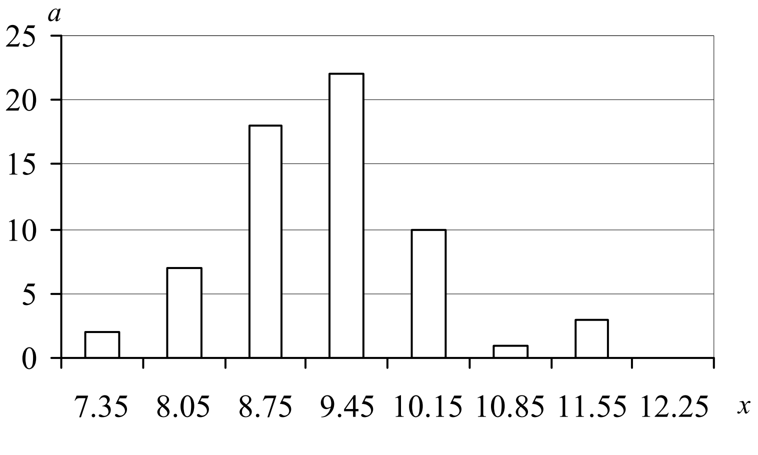


Рис. 3. *Распределение бурозубок по весу тела*

ВЫЧИСЛЕНИЕ ПАРАМЕТРОВ ВЫБОРОК

Средняя арифметическая

Одной из важнейших обобщающих характеристик вариационного ряда является средняя величина признака (обычно обозначается буквой *M*). Существует несколько видов средних (средняя арифметическая – простая и взвешенная, средняя гармоническая, средняя квадратичная), но в практике биологических исследований наибольшее значение имеет средняя арифметическая – величина, вокруг которой «концентрируются» варианты.

Общая формула для определения величины средней арифметической – это отношение суммы значений всех вариант (x_i) выборки к их числу (объему выборки, n):

$$M = \frac{\sum x_i}{n}.$$

В нашем примере с определением массы бурозубок средняя величина равна $M = 9.298412698$ г. При расчетах статистических параметров на ЭВМ следует помнить, что большое количество значащих цифр обычно не имеет никакого биологического смысла. Записывая такие статистические параметры, как средняя и стандартное отклонение, следует оставлять в лучшем случае на одну значащую цифру больше, чем имели значения вариант, а оценки ошибок – на две значащих цифры. Масса тела бурозубок колебалась от 7.3 до 11.9 г, отсюда средняя с учетом округления должна иметь вид $M = 9.3$ г.

Средняя арифметическая характеризует действие только систематических факторов, поскольку сумма случайных отклонений влево и вправо от средней в силу симметричности нормального распределения обращается в нуль. Поэтому модель варианты меняется: $x_i = M \pm x_{случ}$.

В биологических исследованиях зачастую встречается ситуация, когда требуется первичная статистическая обработка большого числа выборок, но необязательно с большой точностью. Это может понадобиться для предварительного рассмотрения и оценки материала, в частности для оперативного выявления общих тенденций его изменчивости, с тем, чтобы в дальнейшем перейти к специальным методам статистического анализа. Для этих случаев предложен простой *экспресс-метод* с использованием полученного для данной выборки размаха значений (*Lim*). В случае нормального распределения средняя арифметическая находится точно по центру (совпадает со значением медианы), т. е. левая и правая границы распределения находятся на одинаковом расстоянии от средней. Исходя из этих соображений, среднюю арифметическую можно рассчитать по формуле медианы:

$$M = Me = \frac{x_{\min} + x_{\max}}{2}.$$

Для бурозубок эта средняя составит $M = (7.3 + 11.9) / 2 = 9.6$ г, что вполне соответствует первой точной оценке.

В случаях, когда необходимо объединить результаты расчетов по нескольким выборкам и на этой основе найти общую среднюю, характеризую-

щую весь изученный материал, пользуются *взвешенной средней*, которая учитывает объемы частных выборок:

$$M = \frac{\sum n_j \cdot M_j}{\sum n_j},$$

где M_j – значение частной средней,

n_j – условные «веса» частного значения, объемы выборок.

Чтобы рассчитать среднюю взвешенную, необходимо значения всех частных средних арифметических помножить на свои «веса», все эти произведения сложить и сумму разделить на сумму весов (общий объем всех выборок). Пусть получены результаты определения средней величины выводка у рыжих полевок (экз. / самку) по месяцам: май 5.0, июнь 5.4, июль 6.2, август 6.0, сентябрь 4.5, причем известно число определений (самок) для каждого месяца: 22, 43, 103, 33 и 5. Взвешенная средняя арифметическая составит:

$M = (5{\cdot}22 + 5.4{\cdot}43 + 6.2{\cdot}103 + 6{\cdot}33 + 4.5{\cdot}5) / (22 + 43 + 103 + 33 + 5) = 5.8.$

Средняя, рассчитанная обычным способом, оказалась заниженной:

$M = (5 + 5.4 + 6.2 + 6 + 4.5) / 5 = 5.4.$

В число прочих констант вариационного ряда входит *медиана* (Me) – значение, делящее размах выборки пополам, и *мода* (Mo) – класс (или значение), представленный наибольшим числом вариант.

Стандартное отклонение

Среднее квадратичное отклонение (или стандартное отклонение) – вторая по значению константа вариационного ряда. Она является мерой разнообразия входящих в группу объектов и показывает, на сколько *в среднем* отклоняются варианты от средней арифметической изучаемой совокупности. Чем сильнее разбросаны варианты вокруг средней, чем чаще встречаются крайние или другие отдаленные классы отклонений от средней вариационного ряда, тем большим оказывается и среднее квадратичное отклонение. Стандартное отклонение есть мера изменчивости признаков, обусловленная влиянием на них случайных факторов. Квадрат стандартного отклонения (S^2) называется *дисперсией.*

Что такое «случайное» при детальном рассмотрении? В формуле модели вариант случайный компонент предстает в виде некой «добавки» к доле варианты, сформированной под действием систематических факторов, $\pm x_{случ.}$. Она, в свою очередь, складывается из эффектов влияния неопределенно большого числа факторов: $x_{случ.} = \sum x_{случ.k}$.

Каждый из этих факторов может обнаружить свое сильное действие (дать большой вклад), а может почти не участвовать в становлении конкретной варианты (слабое действие, незначительный вклад). Причем доля случайной «прибавки» для каждой варианты оказывается различной! Рассматривая, например, размеры дафний, можно увидеть, что одна особь крупнее, другая мельче, поскольку одна родилась на несколько часов раньше, другая – позже, или одна генетически не вполне идентична прочим, а третья росла в

более прогреваемой зоне аквариума и т. д. Если эти частные факторы *не входят в число контролируемых* при сборе вариант, то они, индивидуально проявляясь в разной степени, обеспечивают *случайное* варьирование вариант. Чем больше случайных факторов, чем они сильнее, тем дальше будут разбросаны варианты вокруг средней и тем большим оказывается характеристика варьирования, среднее квадратичное отклонение. В контексте нашей книги термин «случайное» есть синоним слова «неизвестное», «неподконтрольное». Пока мы каким-либо способом не выразим интенсивность фактора (группировкой, градацией, числом), до тех пор он останется фактором, вызывающим случайную изменчивость.

Смысл стандартного отклонения (вариант от средней) выражает формула:

$$S = \sqrt{\frac{\sum (x - M)^2}{(n-1)}},$$

где x – значение признака у каждого объекта в группе,
M – средняя арифметическая признака,
n – число вариант выборки.

Выполнять расчеты удобнее с помощью *рабочей формулы*:

$$S = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{(n-1)}},$$

где $\sum x^2$ – сумма квадратов значений признака для всех вариант,
$\sum x$ – сумма значений признака,
n – объем выборки.

Для примера с массой тела бурозубок стандартное отклонение будет равно: $S = 0.897216496$, а после необходимого округления $S = 0.897$ г.

В некоторых случаях бывает необходимо определить *взвешенное среднее квадратичное отклонение* для суммарного распределения, составленного из нескольких выборок, для которых значения стандартных отклонений уже известны. Эта задача решается с помощью формулы:

$$S_{\Sigma} = \sqrt{\frac{\sum S^2 (n-1)}{\sum n - k}},$$

где S_{Σ} – усредненная величина среднего квадратичного отклонения для суммарного распределения,
S – усредняемые значения стандартного отклонения,
n – объемы отдельных выборок,
k – число усредняемых стандартных отклонений.

Рассмотрим такой пример. Четыре независимых определения веса печени (мг) у землероек-бурозубок в июне, июле, августе и сентябре дали следующие величины стандартных отклонений: 93, 83, 50, 71 (при n = 17, 115, 132, 140). Подставив в вышеприведенную формулу нужные значения, получим стандартные отклонения для суммарной выборки (для всего бесснежного

периода):

$$S_{\Sigma} = \sqrt{\frac{93^2 \cdot 16 + 83^2 \cdot 114 + 503^2 \cdot 131 + 71^2 \cdot 139}{404 - 4}} = 69.9.$$

В случае, если требуется первичная статистическая обработка большого числа выборок, но необязательно с большой точностью, для оценки стандартного отклонения можно воспользоваться *экспресс-методом*, основанным на знании закона нормального распределения. Как уже отмечалось, крайние значения для выборки (с вероятностью $P = 95\%$) можно считать границами, удаленными от средней на расстояние $2S$: $x_{\min} = M - 2S$, $x_{\max} = M + 2S$. Это значит, что в лимите (Lim), в диапазоне от максимального до минимального выборочного значения, укладываются четыре стандартных отклонения:

$Lim = (M + 2S) - (M - 2S) = 4S.$

Однако этот вывод справедлив только по отношению к выборкам большого размера, тогда как для небольших выборок необходимо делать поправки. Рекомендуется следующая формула приблизительного расчета стандартного отклонения (Ашмарин и др., 1975):

$$S = \frac{x_{\max} - x_{\min}}{d},$$

где величина d взята из таблицы 3 (против соответствующего объема выборки, n).

Таблица 3

n	*d*	*n*	*d*	*n*	*d*	*n*	*d*
2	1.128	7	2.704	12	3.258	17	3.588
3	1.693	8	2.847	13	3.336	18	3.640
4	2.059	9	2.970	14	3.407	19	3.689
5	2.326	10	3.079	15	3.472	20	3.735
6	2.534	11	3.173	16	3.532	более	4

Выборочное стандартное отклонение веса тела бурозубок ($n = 63$), рассчитанное по приведенной формуле, составляет:

$S = (11.9 - 7.3) / 4 = 1.15$ г,

что достаточно близко к точному значению, $S = 0.89$ г.

Использование экспресс-оценок стандартного отклонения значительно сокращает время расчетов, существенно не сказываясь на их точности. Отмечается лишь небольшая тенденция к завышению получаемых этим методом значений стандартного отклонения при небольших объемах выборок.

Стандартное отклонение – величина именованная, поэтому с ее помощью можно сравнивать характер варьирования лишь одних и тех же признаков. Чтобы сопоставить изменчивость разнородных признаков, выраженных в различных единицах измерения, а также нивелировать влияние масштаба измерений, используют так называемый *коэффициент вариации (CV)*, безразмерную величину, отношение выборочной оценки S к собственной средней M:

$$CV = \frac{S}{M} \cdot 100\%.$$

В нашем примере с весом тела бурозубок:

$$CV = \frac{S}{M} \cdot 100\% = \frac{0.89}{9.3} \cdot 100\% = 9.6\%.$$

Индивидуальная изменчивость (варьирование) признаков – одна из наиболее емких характеристик биологической популяции, любого биологического процесса или явления. Коэффициент вариации может считаться вполне адекватным и объективным показателем, хорошо отражающим фактическое разнообразие совокупности независимо от абсолютной величины признака. Индекс был создан для унификации показателей изменчивости разных или разноразмерных признаков путем приведения их к одному масштабу. Практика показывает, что для многих биологических признаков наблюдается увеличение изменчивости (стандартного отклонения) с ростом их величины (средней арифметической). При этом коэффициент вариации остается примерно на одном и том же уровне – 8–15%. За увеличение коэффициента вариации ответственны, как правило, растущие отличия распределения признака от нормального закона.

ОСНОВНЫЕ ТИПЫ РАСПРЕДЕЛЕНИЙ ПРИЗНАКОВ

Как уже отмечалось, биометрия изучает случайные события, поведение случайных величин. Начиная биологический эксперимент или приступая к наблюдению, невозможно точно сказать, каков будет результат – уровень численности животных в данном районе, вес еще не отловленных особей, количество сахара в крови через час после введения препарата и т. п. В этом смысле биологические явления случайны, точно не предсказуемы. Однако любому биологу ясно, что случайность эта не абсолютна. Несмотря на сложность точного прогноза, приблизительный результат можно предугадать, в частности, предсказав, что интересующая нас величина будет находиться в пределах некоторого интервала между конкретными минимальными и максимальными значениями. Ясно, например, что рост человека вряд ли превысит два или будет ниже полутора метров. Вариационная статистика может дать и более точный прогноз, ориентируясь на известные законы поведения случайных величин, относящихся к разным типам распределений. При этом под распределением признаков (случайных величин, объектов) понимается соотношение между их значениями и частотой встречаемости.

Среди многих известных типов распределений мы рассмотрим лишь пять (нормальное, биномиальное, Пуассона, альтернативное, полиномиальное, равномерное). Для описания природных явлений иногда реалистичные основания имеет распределение *гипергеометрическое* (безвозвратное изъятие). Распределение *негативное биномиальное* подходит для случая, когда вероятности элементарных событий (p и q) не постоянны.

Распределения *Максвелла* и *Рэлея* имеют умеренную правостороннюю асимметрию и описывают поведение непрерывных положительных случайных величин. Распределения *Парето* и *показательное* пригодны для описания резко правосторонне асимметричных вариационных рядов с перепадом частот. Распределение *логнормальное*, или логарифмически нормальное, характеризуется тем, что логарифмы исходных значений выборки образуют правильное нормальное распределение; эта модель подходит для описания признаков, имеющих распределения с умеренной правосторонней асимметрией, это в первую очередь концентрации веществ в различных средах, т. е. гидрохимические, физиологические и биохимические показатели.

Зная тип распределения, можно воспользоваться разработанными специально для него приемами математической обработки и получить наиболее полную информацию о явлении, точнее оценить различия между параметрами разных выборок.

Нормальное распределение

Наиболее характерный тип распределения *непрерывных случайных величин*, из него можно вывести (к нему сводятся) все остальные. Распределение *симметрично*, причем крайние значения (наибольшие и наименьшие) появляются редко, но чем ближе значения признака к центру (к средней арифметической), тем оно чаще встречается (рис. 4).

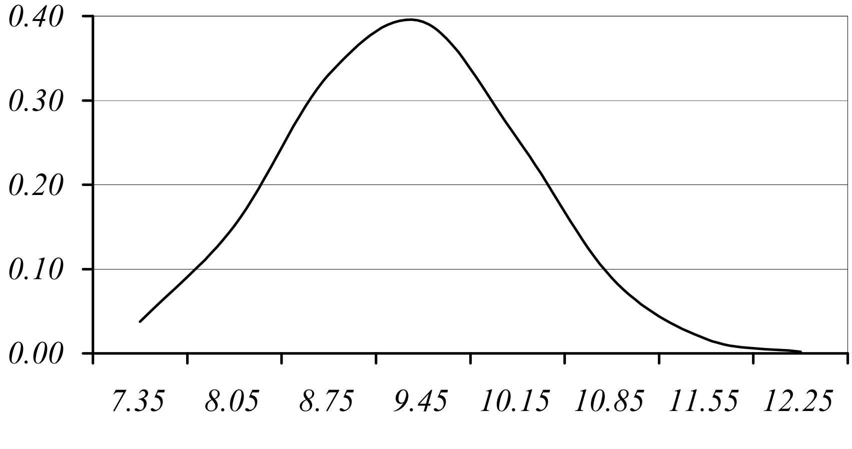


Рис. 4. *Нормальное распределение с параметрами n = 63, M = 9.3, S = 0.79.*

По оси абсцисс – вес тела землероек-бурозубок, по оси ординат – табличные значения для нормального распределения. Рассчитать ординаты нормальной кривой для конкретного значения x_i можно по формуле: $p_i = (1/\sqrt{2\pi}) \cdot e^{-(x_i - M)^2/2 \cdot S^2}$

Среднее квадратичное отклонение примерно 4 раза укладывается в размахе изменчивости признака и по величине значительно *уступает* средней. Геометрически стандартное отклонение равно расстоянию от центра кривой распределения до точки перегиба кривой. Примеры расчета параметров (*M*, *S*) нормального распределения приведены выше.

Биномиальное распределение

Во многом близко к нормальному. Отличие состоит лишь в том, что оно характеризует поведение *дискретных признаков*, *выраженных целыми числами*. Как правило, для описания биологических признаков подходит симметричное биномиальное распределение, у которого дисперсия много меньше средней. Распределение организуется в процессе обора *проб* (объемом больше одного, $m > 1$). Число классов больше двух, $k > 2$.

Примерами описания признаков с помощью биномиального распределения могут служить число поврежденных участков на листьях, число волосков на единице площади шкурки, количество лучей в плавниках рыб, число хвостовых щитков у рептилий, плодовитость (размер выводка) самок и т. п. В основе биномиального распределения лежит альтернативное проявление качественного признака: он может присутствовать у единичного объекта или отсутствовать, проявиться или нет. Отдельный корнеплод может быть больным или здоровым (признак качественный), тогда *проба* из нескольких корнеплодов будет содержать некоторое *число* здоровых корнеплодов (признак количественный), а множество равнообъемных проб образует уже выборку чисел, для которой можно построить гистограмму распределения. Вероятность отдельного события (корнеплод больной) составляет p, а вероятность альтернативного события (корнеплод здоровый) равна $q = 1 - p$. При равенстве вероятностей событий $p = q = 0.5$. большинство проб (вариант) будет иметь около половины возможных событий (поровну больных и здоровых корнеплодов); распределение примет симметричную форму. В случае неравенства вероятностей наблюдается та или иная степень асимметрии распределения.

Рассмотрим результаты изучения плодовитости серебристо-черных лисиц (число щенков на самку) (см. данные на стр. 12). Для построения вариационного ряда берем 8 классов, классовый интервал для этого дискретного признака составит $dx = 1$.

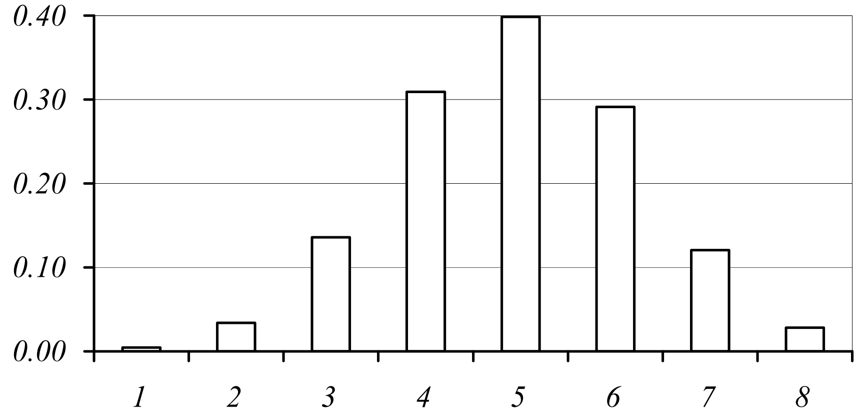


Рис. 5. *Биномиальное распределение (n = 76, M = 4.95, S = 1.33).*

По оси абсцисс – число щенков лисицы на одну самку, по оси ординат – частости (относительные частоты)

Все основные параметры распределения вычисляются по рассмотренным выше формулам:

$$M = \frac{\sum x}{n} = 4.96 \text{ экз./самку},$$

$$S = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{(n-1)}} = 1.33 \text{ экз./самку},$$

$$m_M = \frac{s}{\sqrt{n}} = \frac{1.33}{\sqrt{63}} = 0.1676 \text{ экз./самку},$$

$$m_S = \frac{s}{\sqrt{2 \cdot n}} = \frac{1.33}{\sqrt{2 \cdot 63}} = 0.1185 \text{ экз./самку}.$$

Для расчета параметров биномиального распределения можно воспользоваться другими, более простыми формулами, если предварительно рассчитать вероятности p и q (в нашем случае $p = 0.62$, $q = 0.38$):

$$M = m \cdot p = 8 \cdot 0.62 = 4.96 \text{ экз./самку},$$

$$S = \sqrt{m \cdot p \cdot q} = \sqrt{8 \cdot 0.62 \cdot 0.38} = 1.37 \text{ экз./самку}.$$

Результаты оказываются идентичными с точностью до ошибки округления.

Доверительный интервал для параметров биномиального распределения строится так же, как и для нормального распределения: $M \pm tm_M$, $S \pm tm_S$.

Распределение Пуассона

Это вариант описания стохастического поведения *дискретных количественных признаков* для случаев, когда *вероятность элементарных альтернативных событий неодинакова*, одно из них наблюдается заметно чаще другого ($p << q$) (классический пример – попадание гитлеровских авиационных бомб в разные кварталы Лондона). Закон Пуассона описывает редкие события, происходящие 1, 2, 3 и т. д. раз на сотни и тысячи обычных событий. Поведение биологических объектов, соответствующее закону Пуассона, наблюдается в том случае, когда по пробам случайно распределены редкие объекты. Примеры таких явлений – частота нарушений хромосомного аппарата на каждую тысячу митозов, встречаемость семян сорняка в большой серии навесок семян культурного растения, число повторных попаданий животных в ловушки, встречаемость животных на отрезках длинных маршрутов (или на пробных площадках обширной территории), отловы животных в отдельные промежутки времени при длительных наблюдениях.

Случайная величина, распределенная по закону Пуассона, определяется при подсчете числа элементарных событий *в пробе* (в группе, в навеске, на участке, на этапе). Число объектов в пробе больше 1 ($m > 1$), число классов больше двух ($k > 2$).

Распределение Пуассона резко асимметрично, причем *дисперсия равна средней арифметической*, что может служить критерием для оценки харак-

тера распределения изучаемого признака (рис. 6). В течение одного года (1946) пометили кольцами и выпустили на волю 32 буревестника.

Число повторных отловов, x	Число отловленных животных, a	Число случаев повторного отлова, $x \cdot a$
0	15	0
1	7	7
2	7	14
3	2	6
4	1	4
n	32	31

В последующие пять лет часть из них отлавливали повторно: 7 экз. по одному разу, 7 – по два, 2 – по три, 1 экз. – четыре раза, 15 экз. окольцованных птиц повторно не попадались. Число классов составляет k = 4, интервал dx = 1. Асимметрия в частотах встречаемости птиц позволяет предполагать распределение Пуассона.

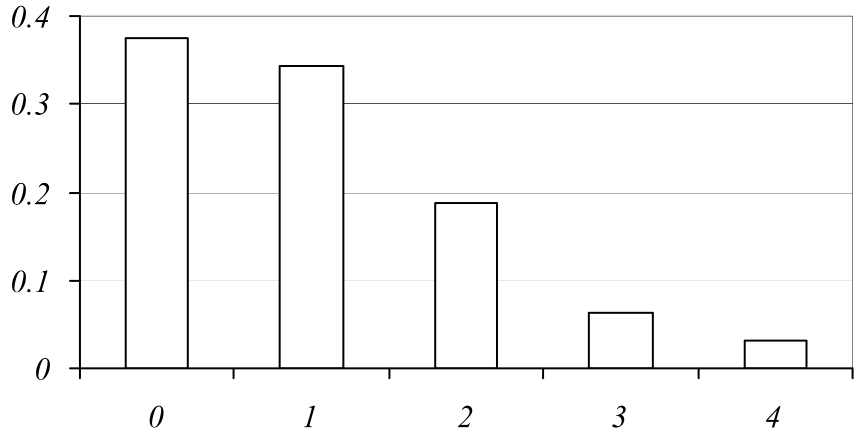


Рис. 6. *Распределение Пуассона с параметрами n = 32, M ≈ S² = 0.968.*

По оси абсцисс – число повторных отловов, по оси ординат – частости (относительные частоты)

Расчеты показали, что средняя арифметическая (M) примерно равна дисперсии (S^2):

$$M = \frac{\sum x}{n} = \frac{31}{32} = m \cdot p = 4 \cdot 0.242 = 0.968 \text{ экз.},$$

$$S = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{(n-1)}} = \sqrt{\frac{69 - \frac{(32)^2}{32}}{(32-1)}} = 1.121 \text{ экз.}, S^2 = 1.257,$$

$S^2 \approx M$.

Критерий Фишера не выявил достоверных отличий между средней и

дисперсией: $F = 1.257 / 0.968 = 1.157 < F_{(0.05,31,31)} = 1.8$, что свидетельствует о соответствии наблюдаемого распределения закону Пуассона.

Возможен расчет параметров по более простым формулам:

$M = m \cdot p,\ S = \sqrt{m \cdot p}$.

Доверительный интервал для параметров распределения Пуассона определить несколько сложнее, чем для других типов (Ивантер, Коросов, 2003).

Альтернативное распределение

Распределение *дискретной случайной величины*, имеющей лишь два противоположных (разнокачественных) значения (два класса, $k = 2$). В одной пробе (в одном наблюдении) содержится одна варианта ($m = 1$), одно из двух возможных значений. Вероятности каждого из них могут быть равны ($p = q$) либо не равны ($p < q;\ p > q$). Примеры: самцы и самки, больные и здоровые организмы, сработавшие и пустые ловушки на одной учетной линии, два варианта аллельных признаков, вакцинированные и невакцинированные пациенты среди заболевших и др. (рис. 7). Вычисления констант достаточно просты и не требуют построения вариационного ряда.

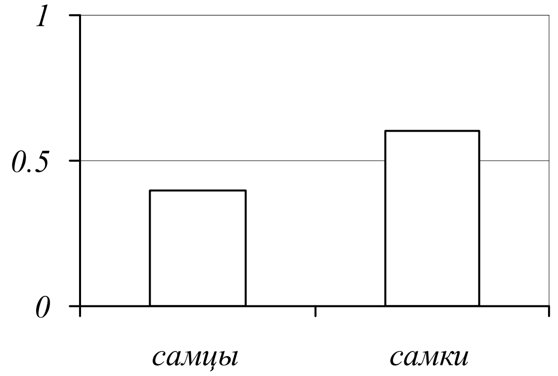


Рис. 7. *Альтернативное распределение (два класса вариант). По оси ординат – частости (доли) этих групп*

Важнейшей характеристикой является доля (p) вариант определенного вида (А), представленных общим числом n_A в пределах выборки объемом n:

$$p = \frac{n_A}{n}.$$

Если исходы отдельных испытаний выразить числами 0 или 1 (что аналогично отбору проб с объемом $m = 1$), доля вариант совпадает со средней арифметической, рассчитанной для всех значений:

$$M = \frac{\sum x}{n}.$$

Например, результат отловов полевок из природных популяций показал, что в исследуемой группе (200 особей) было 120 самок и 80 самцов. В данном случае мы имеем дело с альтернативным признаком (самка – самец). Из 200 проб 120 содержат самок (значение 1), 80 – не содержат (зна-

чение 0), так получаем выборку n = 200:

0000000000000000000000000000000000000000000000000000000000000000001
1111111111111111111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111111.

Доля вариант со значением 1 составляет:

$$p = \frac{n_A}{n} = \frac{120}{200} = 0.6,$$

что совпадает со средней арифметической для всего ряда:

$$M = \frac{\sum x}{n} = \frac{120}{200} = 0.6.$$

Для альтернативного распределения могут применяться те же формулы расчета выборочных параметров, что и для биномиального распределения. Средняя (доля самок):

$M = m \cdot p = 1 \cdot 0.6 = p = 0.6.$

Стандартное отклонение (при m = 1):

$S = \sqrt{m \cdot p \cdot q} = \sqrt{p \cdot q} = \sqrt{0.6 \cdot 0.4} =$ 0.24.

Ошибка средней (ошибка доли самок):

$$m = \frac{S}{\sqrt{n}} = \frac{0.24}{\sqrt{200}} = 0.017.$$

Доверительный интервал для альтернативных признаков (их долей, процентов и частот) строится с помощью φ-преобразования Фишера, что дает более точные границы, особенно если доли сильно отличаются. Сначала вместо значения доли (процента) одного признака объектов берут значение φ (фи), найденное по формуле $\varphi = 2 \cdot \arcsin\sqrt{p}$ или по таблице 10*П*. Затем вычисляют ошибку: $m_\varphi = 1/\sqrt{n}$, обе доверительные границы $\varphi_{лев.} = \varphi - tm_\varphi$, $\varphi_{прав.} = \varphi + tm_\varphi$, после чего с помощью таблицы 10*П* переводят найденные значения обратно в проценты.

Найдем доверительные границы для доли самок полевок $p = 0.6$ при уровне значимости $\alpha = 0.05$. Используя таблицу 10*П* и проводя расчеты, получаем: φ(60%) = 1.772, $m_\varphi = 1/\sqrt{200}$ = 0.0707,

$\varphi_{лев.}$ = 1.772 − 1.96 · 0.0707 = 1.6334, $\varphi_{прав.}$ = 1.772 + 1.96 · 0.0707 = 1.9106, $p_{лев.}$(1.6334) = 53.1%, $p_{прав.}$(1.9106) = 66.4%.

Доля самок в генеральной совокупности (популяции полевок) составляет минимум 53.1%, максимум 66.4%.

Полиномиальное распределение

Наблюдается для *качественных признаков*, имеющих не два альтернативных свойства, но *несколько возможных проявлений качества*. Примеры полиморфизма популяций – из этой области. В их числе варианты окраски покровов и волос, типы рисунков в определенных областях тела, способы жилкования листьев растений или крыльев насекомых, варианты расположения и формы щитков рептилий и другие проявления множественности фено-

типов особей. Формализуя описание, укажем, что в одной пробе содержится одна варианта ($m = 1$), но типов вариант (морф, фенотипов) больше, чем два ($k > 2$).

Примером полиномиального (иначе – мультиномиального) распределения может служить встречаемость 4 фенов головы живородящей ящерицы – 4 вариантов контакта лобно-носового, предлобных и лобного щитков (рис. 8). Лучше всего выборка может быть представлена вариационным рядом – частотами (p_j) встречаемости в популяции особей с данным (j-м) проявлением качественного признака и общим числом морф (k). Для более емкого представления ряда и учета характера распределения частот между разными морфами используется величина «среднее число фенотипов»: $\mu = \Sigma(p_j)^2$, статистическая ошибка которой рассчитывается так: $m_\mu = \sqrt{\dfrac{\mu \cdot (k - \mu)}{n}}$.

Среднее число фенотипов (μ) равно числу фенотипов (k) только тогда, когда частоты всех фенотипов одинаковы ($p_1 = p_2 = \ldots = p_j \ldots = p_k$), и меньше во всех других случаях.

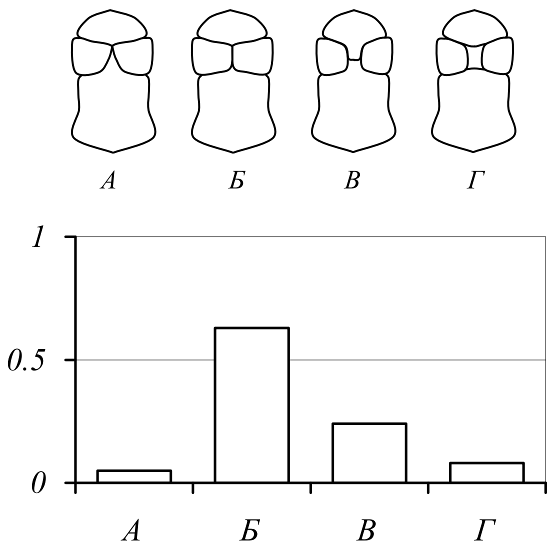


Рис. 8. *Полиномиальное распределение (4 фена головы ящерицы).*

По оси ординат – частости фенов среди 64 сеголетков живородящей ящерицы, отловленных под Петрозаводском

Равномерное распределение

Частный случай распределения альтернативного и полиномиального. Равномерное распределение характеризуется одинаковой частотой встречаемости всех значений дискретного признака ($p = q$ для двух классов или $p_1 = p_2 = \ldots = p_j \ldots = p_k$ для нескольких классов). Такой тип распределения можно использовать для формулирования гипотез при анализе частот генов и фенов в популяциях, при подсчете тест-организмов, выживших в токсикометрическом эксперименте, можно предположить, что ветви дерева могут равномерно располагаться по сторонам света.

СТАТИСТИЧЕСКАЯ ОЦЕНКА ГЕНЕРАЛЬНЫХ ПАРАМЕТРОВ

Биометрия изучает поведение биологических случайных величин, которые точно не предсказуемы, хотя и не абсолютно случайны. В этом разделе будут рассмотрены способы определения диапазона возможной изменчивости изучаемых биологических признаков. Приблизительный прогноз всегда можно дать в виде интервала между конкретными минимальными и максимальными значениями, в пределах которого будет находиться интересующая нас величина. Ясно, например, что рост очередного встречного взрослого человека вряд ли превысит два метра или будет меньше полутора метров. Более точный (вероятностный) прогноз можно дать, ориентируясь на распределение случайных величин. *Распределение – это соотношение между значениями случайной величины и частотой их встречаемости.* Как мы видели на примере веса тела землероек, числовые значения вариант располагаются в некоторой ограниченной зоне, в центре которой их особенно много, а по краям мало. Ключом к получению вероятностного прогноза служит знание законов распределения случайных величин. Очень большое число случайных величин, распространенных в природе, может быть описано с помощью закона нормального распределения, который задается уравнением:

$$p = \frac{1}{\sqrt{2\pi}} \cdot e^{-t^2/2},$$

где $t = \frac{(x-M)^2}{S}$ – нормированное отклонение;

M, S – параметры нормального распределения.

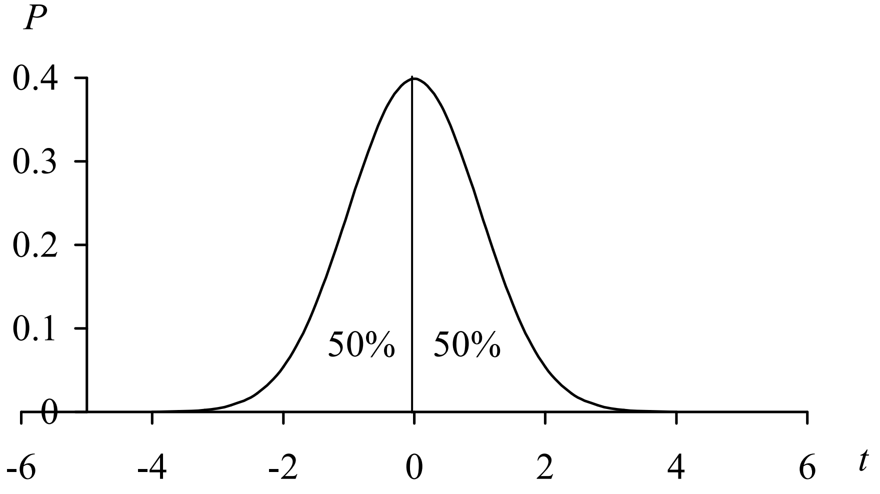
Эта модель лежит в основе многих статистических методов.

Свойства нормального распределения

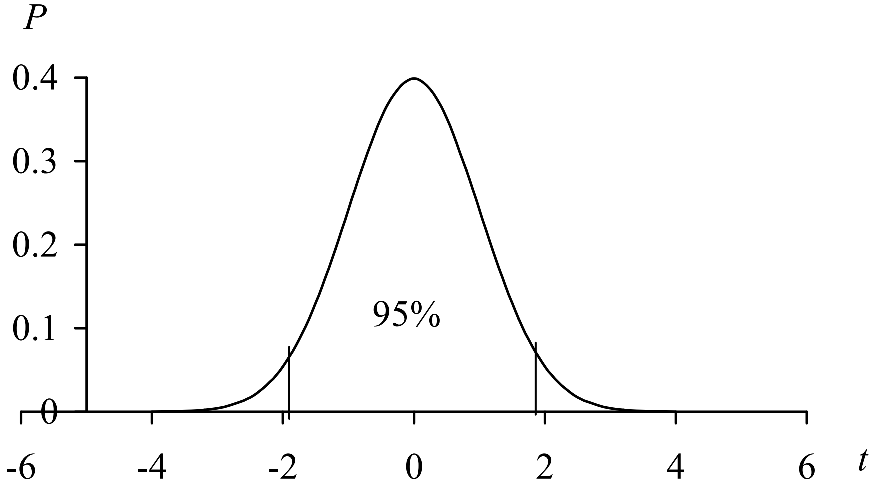
Приведенное уравнение определяет ход кривой линии, имеющей характерную колоколообразную форму, и позволяет вычислить *ординаты нормальной кривой*, или «плотность вероятности» (p). *Вероятность (статистическая, или частость) – численная мера возможного, определяется как отношение числа вариант (исходов испытаний) определенного вида к общему числу вариант (опытов).* Поскольку нормальное распределение характерно для непрерывных случайных величин, говорят не о вероятности какого-то определенного значения варианты, но о «*плотности вероятности*», отражая тем самым плавность изменения вероятности значений для разных значений t, чем ближе к центру распределения, тем плотность вероятности выше.

С помощью представленного выше уравнения можно рассчитать вероятность появления нового значения случайной величины t в интервале той или иной ширины и дать *статистическую оценку – найти интервал значений признака, в котором с той или иной вероятностью заключено значение генерального параметра*. Формула количественно выражает вполне определенные свойства поведения случайной величины, из которых можно назвать следующие практически важные следствия:

1. Все варианты лежат в интервале плюс-минус бесконечность. Иными словами, с вероятностью $P = 1$ ($P = 100\%$) мы вправе ожидать появление новой варианты в пределах от $-\infty$ до $+\infty$. Слева и справа от средней арифметической лежит по 50% вариант (свойство симметрии нормального распределения), т. е. с вероятностью $P = 0.5$ (50%) можно предсказать появление новой варианты в интервалах $M - \infty$ и $M + \infty$.



2. Между $M - 1.96S$ и $M + 1.96S$ лежит 95% вариант. Это позволяет с 95%-ой вероятностью предполагать, что новая варианта окажется в интервале $M \pm 1.96S$ (округленно $M \pm 2S$ – так называемое *правило двух стандартных отклонений*).



3. С вероятностью $P = 0.99$ значение новой варианты будет заключено в пределах $M \pm 2.58S$ и с вероятностью $P = 0.999$ – в интервале $M \pm 3.3S$.

Исходя из сказанного можно оценить вероятность появления новых значений признака. В отношении непрерывных случайных величин (метрических признаков) эта процедура сводится к так называемой *интервальной оценке*. Для полученных ранее характеристик, массы бурозубок, средней $M = 9.26$ и стандартного отклонения $S = 0.79$ (г), находим прогнозный интер-

вал: $M \pm 1.96S = 9.26 \pm 1.53$. Новое значение признака с вероятностью $P = 0.95$ между 7.68 и 10.82 г. Предсказание веса землероек, конечно, не имеет большого практического значения. Гораздо важнее может быть прогноз численности ценных промысловых видов, сельскохозяйственных вредителей, вспышек болезней, урожая культурных растений и т. п.

Важнейшее значение для практического применения имеет «соглашение о 95%». В соответствии с ним совокупности, состоящей из 95% особей (объектов), мы доверяем так же, как и 100%-й. Термин «*доверительная вероятность* $P = 0.95$» означает, что, согласно принятому допущению, *95% вариант достаточно полно характеризуют изучаемое явление* (в данном случае изменчивость веса землероек), что позволяет ограничиться рассмотрением вариант в области $M \pm 1.96S$, охватывающей эту 95%-ю совокупность. Так, мы принимаем, что нормальный вес землероек данного вида может изменяться в пределах 7.7–10.8 г, не больше и не меньше. За этими пределами мы обнаруживаем животных иного вида или статуса.

При этом в биометрии обычно довольствуются доверительной вероятностью $P = 0.95$ (уровень значимости $\alpha = 0.05$), хотя в наиболее ответственных исследованиях принимают и более строгие уровни – $P = 0.99$ и $P = 0.999$. Однако это имеет смысл лишь при очень больших выборках исходных данных, точно описывающих закономерности изменчивости признаков. Обычно же выборки не очень велики, что позволяет ограничиться меньшей степенью доверительной вероятности $P = 0.95$.

Уровень значимости – понятие, альтернативное доверительной вероятности ($\alpha = 1 - P$). Для доверительной вероятности 0.95 уровень значимости составляет 0.05, а для 0.99 и 0.999 – соответственно 0.01 и 0.001. Уровень значимости, равный 0.05 (5%), можно интерпретировать так: имеется всего 5% шансов, что полученная величина не будет соответствовать изучаемой совокупности. *Уровень значимости – это тот теоретический процент значений нормального распределения, который можно отбросить, не учитывать, дабы с меньшими усилиями получить основную информацию об изучаемом явлении.* Можно целую жизнь положить на попытки отловить обыкновенную землеройку-бурозубку весом 2.5 г, но так и не собрать выборку, достаточную по объему, чтобы это реализовать (миллионы особей). Для практического понимания достаточно знать, что уровень значимости – это вероятность ожидаемой ошибки наших выводов, *вероятность того, что данный статистический вывод не верен*. И с этой позиции 5% – достаточно мало. Использование доверительной вероятности и уровня значимости можно назвать теоретической базой разумного ограничения времени и масштабов исследования, позволяющей получить достоверную общую информацию за счет исключения ничтожной доли частной.

Генеральная совокупность

Генеральная совокупность – все варианты одного типа. В предметной биологии это понятие можно интерпретировать как мыслимое множество вариант, сформированных при одинаковых (внешних и внутренних) условиях.

Теоретическая бесконечность генеральной совокупности означает, что ее никогда нельзя познать до конца, в действительности мы всегда имеем дело с выборками. *Выборочная совокупность, выборка – это множество вариант одного типа, ограниченное способом отбора* (методами получения вариант) *из генеральной совокупности*. Отличие выборок от генеральной совокупности состоит в том, что действующие в генеральной совокупности факторы не могут проявиться *в полной мере* в любой отдельной выборке. Каждая новая выборка обязательно будет отличаться от предыдущей *в силу случайности*, варианты новой выборки будут нести одинаковый отпечаток действия доминирующих факторов, но разные следы действия случайных факторов. По этой причине параметры (средняя M и стандартное отклонение S) разных выборок из одной генеральной совокупности никогда не совпадут ни друг с другом, ни со значениями генеральных параметров (обычно обозначаемых буквами μ, σ), они будут немного отличаться, смещаясь относительно друг друга и варьируя вокруг генеральных значений.

Отличие генеральных параметров от их оценок по выборкам состоит еще и в том, что в первом случае они рассчитаны по всем вариантам, а во втором – по ограниченному их числу. Интуитивно понятно, что чем меньше объем выборок, тем менее точным будут выборочные оценки генеральных параметров, и, напротив, чем больше выборка, тем ближе выборочные средние и дисперсии лежат к генеральным значениям. Это явление называется *законом больших чисел – с ростом числа наблюдений значения выборочных параметров стремятся воспроизвести генеральные*.

Ошибка репрезентативности выборочных параметров

По части никогда не удается полностью охарактеризовать целое, всегда остается вероятность того, что выборочная оценка недостаточно близка к значению параметра генеральной совокупности, имеет некоторую ошибку. *Отличия значений выборочных параметров от генеральных называются ошибкой репрезентативности данного параметра*, или просто (статистической) ошибкой. При увеличении объема выборки ошибки репрезентативности стремятся к нулю (следствие закона больших чисел). Численно выраженные статистические ошибки служат мерой тех пределов, в которых выборочные параметры могут отклоняться от значений генеральных параметров. Например, если для нескольких выборок, полученных из одной и той же генеральной совокупности, посчитать средние, а затем оценить изменчивость этих средних, то стандартное отклонение средних (S_M) и будет численной мерой ошибки репрезентативности выборочной средней. Она обозначается буквой m.

Величина ошибки тем больше, чем больше варьирование признака (S) и чем меньше выборка (n). Ошибку репрезентативности имеют все статистические параметры, рассчитанные по выборке. Для практики статистического оценивания разработаны специальные формулы. Для нормального распреде-

ления они имеют следующий вид. Ошибка средней: $m_M = \frac{S}{\sqrt{n}}$,

ошибка стандартного отклонения: $m_S = \frac{S}{\sqrt{2 \cdot n}}$,

ошибка коэффициента вариации: $m_{CV} = \frac{CV}{\sqrt{2 \cdot n}}$.

Вычисленные значения ошибок подставляют к соответствующим параметрам со знаками плюс-минус (параметр ± ошибка) и в такой форме представляют в научных отчетах и публикациях. В примере с бурозубками для разных параметров имеем:

$$m_M = \frac{0.89}{\sqrt{63}} = 0.113039, \text{M} = 9.3 \pm 0.11 \text{ г},$$

$$m_S = \frac{0.89}{\sqrt{2 \cdot 63}} = 0.07993, \text{S} = 0.89 \pm 0.079 \text{ г},$$

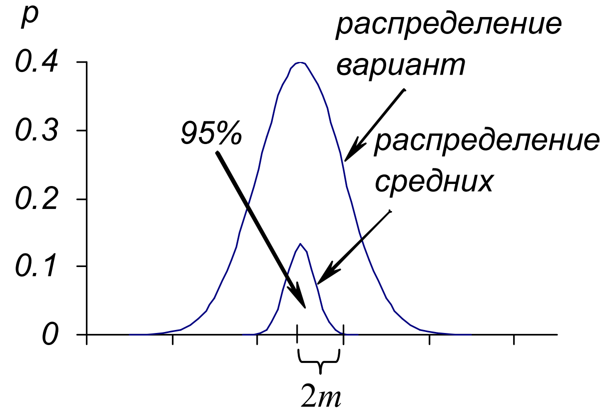
$$m_{CV} = \frac{9.6}{\sqrt{2 \cdot 63}} = 0.859613, \text{CV} = 9.6 \pm 0.9\%.$$

Не следует путать статистическую ошибку с методическими ошибками и ошибками точности (точности измерений, анализов, подсчетов и т. д.), хотя методические погрешности и увеличивают ошибку репрезентативности, но другим путем – методические огрехи увеличивают изменчивость признака, стандартное отклонение.

При всей неизбежности статистической ошибки она может быть сведена к минимуму отбором достаточного числа особей (вариант). С ростом объема выборки оценки параметров стабилизируются, а их ошибки репрезентативности уменьшаются.

Доверительный интервал

Параметры генеральной совокупности практически всегда остаются неизвестными, о них судят по выборочным оценкам, используя для этого значения ошибок репрезентативности. Теоретические исследования поведения выборочных средних (как случайных величин) показали, что они подчиняются нормальному закону, большинство из них (95%) находится поблизости от генеральной средней – в диапазоне $M_{ген.} \pm 1.96m$ (приближенно $\pm 2m$). Это обстоятельство позволяет делать обратное заключение – генеральная средняя находится в диапазоне $M_{выбор.} \pm 1.96m$, т. е. предсказывать ширину интервала, в котором заключен генеральный параметр, давать *интервальную оценку* генеральному параметру.



В соответствии с законом нормального распределения можно ожидать, что генеральный параметр (истинное значение) окажется в интервале

от $M - tm$ до $M + tm$,

где m – ошибка средней арифметической,

t – квантиль распределения Стьюдента (табл. 6*П*) при данном числе степеней свободы (df) и уровне значимости (обычно $\alpha = 0.05$).

Сказанное можно перефразировать так: с вероятностью P можно ожидать, что генеральная средняя находится в доверительном интервале $M \pm tm$, построенном вокруг выборочной средней арифметической M.

Доверительный интервал – интервал значений изучаемого признака, в котором с той или иной вероятностью P находится значение генерального параметра.

Возвращаясь к примеру о весе землероек-бурозубок, мы теперь можем записать доверительные интервалы при разных уровнях вероятности (граничные значения t взяты для случая $n = \infty$):

для $P = 0.95$ $M \pm tm = 9.3 \pm 1.96 \cdot 0.11 = 9.3 \pm 0.21$ г;

для $P = 0.99$ $M \pm tm = 9.3 \pm 2.58 \cdot 0.11 = 9.3 \pm 0.28$ г.

Здесь искомая генеральная средняя величина веса землероек с вероятностью $P = 95\%$ находится в пределах 9.11-9.53 г, а при $P = 99\%$ – 9.04–9.6 г.

Если объем выборки, для которой были получены параметры и ошибка репрезентативности m, был невелик ($n < 50$), то необходимо вводить поправки на объем выборки, расширяя область возможного пребывания генерального параметра. Это понятно, поскольку при дефиците информации любые заключения не могут быть очень точными. Так, для выборки объемом $n = 20$ экз. ошибка средней составит $m_M = \frac{0.89}{\sqrt{20}} = 0.19901$ г, а доверительный интервал $M \pm tm = 9.3 \pm 2.09 \cdot 0.2 = 9.3 \pm 0.41$ г – от 8.9 до 9.7 г (при уровне значимости $\alpha = 0.05$ и числе степеней свободы $df = n - 1 = 20 - 1 = 19$ табличная величина статистики Стьюдента равна $t = 2.09$).

Аналогичным образом можно построить доверительный интервал для стандартного отклонения ($S \pm tm_S$), коэффициента вариации ($CV \pm tm_{CV}$), а также других статистических параметров (коэффициентов асимметрии, эксцесса, регрессии, корреляции).

Определение точности опыта

В практике биометрического анализа используется относительная ошибка измерений – «показатель точности опыта» – отношение ошибки средней к самой средней арифметической, выраженное в процентах: $\varepsilon = \frac{m}{M} \cdot 100\%$. Чем точнее определена средняя, тем меньше будет ε, и наоборот. Точность считается хорошей, если ε меньше 3%, и удовлетворительной при $3 < \varepsilon < 5\%$. Иначе приходится собирать дополнительный материал. В примере показатель точности составил $\varepsilon = (0.11 / 9.3) \cdot 100 = 1.2\%$, что говорит о достаточной надежности выборочной оценки.

Оптимальный объем выборки

В биологических исследованиях часто заранее требуется установить число наблюдений, достаточное для получения репрезентативных оценок генеральной совокупности.

Для непрерывных признаков метод состоит в том, чтобы, используя известные соотношения между средней, стандартным отклонением, ошибкой средней, плотностью вероятности распределения Стьюдента, найти число степеней свободы, соответствующее доверительному интервалу для средней при уровне значимости $\alpha = 0.05$. Объем выборки, достаточной для получения результата заданной точности, находят по формуле:

$$n = \left(\frac{t \cdot CV}{\varepsilon} \right)^2,$$

где n – объем выборки,

t – граничное значение из таблицы распределения Стьюдента (табл. 6*П*), соответствующее принятому уровню значимости при планируемом объеме выборки,

CV – приблизительное значение коэффициента вариации (%),

ε – планируемая точность оценки (погрешности) (%).

Рассчитаем необходимый объем условной выборки, обеспечивающий хорошую точность $\varepsilon = 3\%$, для уровня значимости $\alpha = 0.05$ ($t = 1.98$, для $df \approx 100$) и для коэффициента вариации $CV = 12\%$ (такова относительная изменчивость многих размерно-весовых признаков животных):

$$n = \left(\frac{1.98 \cdot 12}{3} \right)^2 = 62.726 \approx 63 \text{ экз.}$$

Если исследуется фенотипическое (видовое) разнообразие (дискретный признак), может возникнуть задача определения минимального объема выборки, в которой будет присутствовать хотя бы один экземпляр с определенным фенотипом (Животовский, 1991). С позиций теории вероятности задача ставится так: определить объем выборки, в которой с вероятностью P можно ожидать присутствие особи с признаком, частота которого в генеральной совокупности составляет π. Предлагается следующая формула:

$$N = \frac{\ln(1-P)}{\ln(1-\pi)}.$$

В первом приближении значение π можно определить приблизительно по имеющимся данным. Что же касается вероятности P, то ее уровень довольно сильно влияет на величину необходимого объема выборки. Для большей надежности следует брать $P = 0.99$, но тогда возрастет объем работ; не столь высокие требования ($P = 0.95$) могут и не позволить найти искомый фенотип. В частности, при уровне вероятности $P = 0.95$ и предположительной частоте фенотипа в популяции $\pi = 0.05$ потребуется

$$N = \frac{\ln(1-0.95)}{\ln(1-0.05)} = 58.4 \approx 59 \text{ экз.},$$

чтобы отловить хотя бы одну особь с этим дискретным признаком.

ОЦЕНКА ПРИНАДЛЕЖНОСТИ ВАРИАНТЫ К ВЫБОРКЕ

Иногда встречается ситуация, когда одна из полученных вариант сильно отличается от остальных. Можно ли такие резко выделяющиеся значения использовать при дальнейших расчетах? В терминах математической статистики поставленный вопрос звучит так: *относится ли данная варианта вместе с другими вариантами изучаемой выборки к одной и той же генеральной совокупности или к разным?* Его можно сформулировать и по-другому: сформировано ли данное значение варианты под действием тех же доминирующих и случайных факторов, что и все остальные варианты данной выборки, или это были иные факторы? Здесь возможны два ответа.

1. Факторы те же, т. е. все варианты взяты из одной и той же генеральной совокупности.

2. Факторы иные, т. е. особенная варианта и выборка порознь взяты из разных генеральных совокупностей.

Ответ на этот вопрос можно получить с использованием рассмотренных выше свойств нормального распределения. Так, если все варианты были взяты из одной генеральной совокупности, значит, они должны отличаться друг от друга только в силу случайных причин и (с вероятностью $P = 0.95$) находиться в диапазоне $M \pm 2 \cdot S$. Иными словами, по случайным причинам варианты достаточно большой выборки будут отклоняться влево или вправо от средней не более чем на $2 \cdot S$: $x - M < 2 \cdot S$ или $(x - M)/S < 2$.

Эта величина, *нормированное отклонение*, и служит безразмерной характеристикой отклонения отдельной варианты от средней арифметической:

$$t = \frac{x - M}{S} \sim t_{табл.},$$

где t – критерий выпада (исключения),

x – выделяющееся значение признака,

M – средняя величина для группы вариант,

$t_{табл.}$ – стандартные значения критерия выпадов, определяемые свойствами нормального распределения, их можно найти по табл. *5П* для трех уровней вероятности (для больших выборок обычно пользуются значением $t_{табл.} = 2$ при $P = 0.95$, или $\alpha = 0.05$).

Для вариант, принадлежащих к изучаемой достаточно большой выборке, нормированное отклонение меньше двух (с вероятностью $P = 0.95$): $t < 2$. В случае действия на варианту некоего необычного фактора, она окажется за пределами указанного диапазона $M \pm 2S$ и ее нормированное отклонение будет равно или больше двух: $t \geq 2$.

Нормированное отклонение есть простейший *статистический критерий*, который помогает определять так называемые «выскакивающие» варианты и решать вопрос о возможности их отбрасывания как артефактов (исключать из дальнейшей обработки). После такой «чистки» параметры выборки должны быть рассчитаны заново. К оценке чужеродности вариант, как и к другим методам статистики, нельзя подходить формально; цель биометриче-

ского исследования всегда состоит в том, чтобы понять специфику явления. В частности, «отскакивающая» варианта может быть следствием того, что признак имеет иное, *не*-нормальное распределение.

Рассмотрим работу критерия на примере. При измерении длины черепа взрослых самцов обыкновенной землеройки-бурозубки получены выборки с такими параметрами: $M = 18.8$, $S = 0.3$ мм. Общее число животных $n = 85$. Среди прочих вариант два больших значения (19.2 и 21.0) вызывали сомнения. Определим для них критерии выпада:

$$t_1 = \frac{19.2 - 18.8}{0.3} = 1.3 < 2\,,\; t_2 = \frac{21.0 - 18.8}{0.3} = 7.3 > 2\,.$$

Согласно таблице *5П*, критическое значение нормированного отклонения для уровня значимости $\alpha = 0.05$ и $n = 85$ равно $t = 2.0$. Поскольку первое полученное значение (1.3) меньше табличного (2), первый из сомнительных результатов исключать не следует, а второй должен быть отброшен – критерий выпада (7.3) превышает табличное значение (2).

Понятие нормированного отклонения позволяет ввести важнейшее понятие статистики. *Статистика – безразмерная случайная величина, которая имеет известный закон распределения и используется в качестве критерия для проверки статистических гипотез.*

В этом смысле нормированное отклонение есть статистика. Во-первых, это безразмерная величина, поскольку единицы измерения числителя ($x_i - M$) и знаменателя (S) взаимно уничтожаются. Во-вторых, нормированное отклонение имеет вполне определенное распределение (в случае непрерывных признаков – нормальное) со своими параметрами (рис. 9). Его средняя равна нулю $M_t = t_M = (M - M) / S = 0$, а стандартное отклонение равно единице $S_t = t_S = (S - M) / S = (S - 0) / S = S / S = 1$.

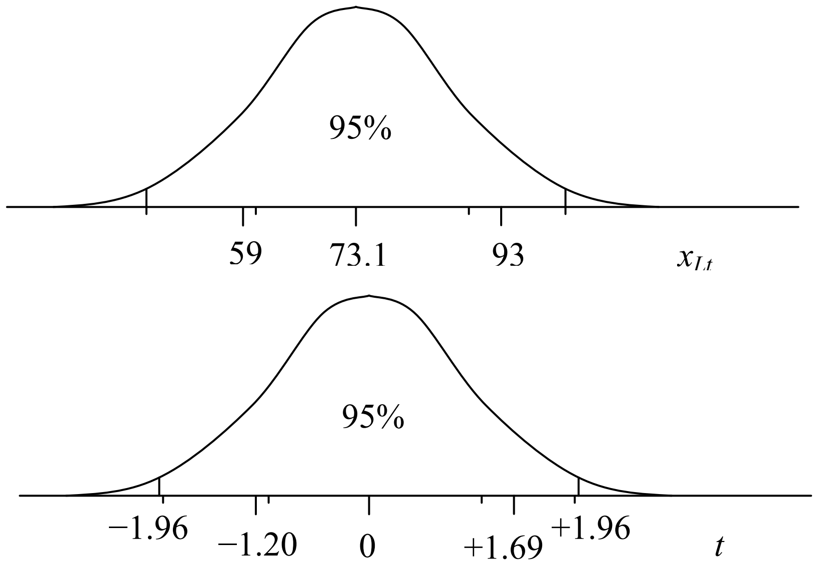


Рис. 9. *Переход от реального признака x к нормированному отклонению t*

Нормированное отклонение – универсальная величина. Какой бы признак (имеющий нормальное распределение) мы ни брали, его значения можно выразить в виде расстояния от центра в единицах стандартного отклонения, т. е. на сколько S данное значение x отклонилось от M. При этом, как следует из свойств нормального распределения, крайние значения в 95% случаев не будут принимать значения меньше −2 и больше 2.

С помощью нормированного отклонения можно, например, оценивать отличия разнокачественных объектов (пород и сортов, видов, популяций, генераций и пр.), причем даже по разным признакам.

Нормированное отклонение можно использовать и для сравнительной оценки разных индивидов по одному и тому же признаку. Например, если сопоставляемые по относительному весу сердца молодая и взрослая землеройки-бурозубки демонстрируют одинаковые показатели (10.5 мг%), то это, тем не менее, не означает их сходства по изучаемому признаку. Используя известную информацию (у молодых средний индекс сердца равен $M = 10.0$ при стандартном отклонении $S = 1.3$, у взрослых – $M = 11.8$, $S = 1.1$), рассчитаем нормированное отклонение для молодого зверька $t_1 = \frac{10.5 - 10}{1.3} = 0.3$ и для взрослого $t_2 = \frac{10.5 - 11.8}{1.1} = -1.2$. Налицо существенное различие: взрослый зверек имеет относительно низкий показатель сердечного индекса, а молодой близок по этому признаку к видовой норме.

Наибольшее развитие такой подход получает в процедурах обработки многомерных данных, при исследовании объектов, охарактеризованных по многим признакам, методом корреляций, главных компонент, при их кластеризации и т. п. Во многих случаях обработка многомерного массива начинается с *нормирования* данных по формуле нормированного отклонения.

ОЦЕНКА РАЗЛИЧИЙ ДВУХ ВЫБОРОК

В любых биологических экспериментах и наблюдениях особое значение имеют различия, на основании которых судят об эффективности действия тех или иных факторов, например, по разности между опытной и контрольной группами делают заключение о результатах опыта. Точно так же по соответствующим изменениям морфофизиологических показателей определяют возрастные, сезонные и популяционные особенности животных. При этом особенно важно оценить статистическую *достоверность разности*, т. е. определить, можно ли данное различие считать закономерным, *характерным для всей генеральной совокупности* и рассматривать его как результат действия особенных факторов, или же оно случайно и является следствием недостаточного количества данных и в следующих опытах может не проявиться.

Обнаружение достоверных отличий статистических параметров – первый шаг к познанию новых биологических закономерностей, причем количественно доказанных. Ответ на вопрос о достоверности или случайности отличий дают статистические критерии, среди которых самые распространенные критерии t Стьюдента и F Фишера. Вычисление их ведется по специальным формулам (различным в зависимости от сравниваемых параметров и типов распределения). Полученные этим способом значения критериев (для чего в формулы подставляются экспериментальные данные) сравнивают с табличными при выбранном уровне значимости (обычно 0.05) и числе степеней свободы (объемы выборок без числа ограничений). Результатом такого сравнения должен стать один из двух вариантов следующего статистического вывода. Если полученное значение (величина) критерия больше табличного, значит, различия между параметрами при заданном уровне значимости и установленном числе степеней свободы достоверны, в разных выборках действительно проявилось действие разных факторов или разных уровней одного фактора. Если же полученная величина критерия меньше табличной, то при данном уровне значимости и числе степеней свободы различия между параметрами недостоверны. Последнее говорит о том, что различия случайны, никакого определенного вывода о побудительных причинах отличий сделать нельзя, нулевая гипотеза остается неопровергнутой.

При сравнении выборок по степени выраженности признака говорят о достоверности (недостоверности) отличий средних арифметических и долей, а при сравнении по уровню изменчивости показателей – о достоверности (недостоверности) отличий стандартных отклонений (дисперсий) и коэффициентов вариации. Особый случай представляет сравнение двух выборок по характеру распределения (достоверность отличия частот), а также общее отличие выборок без указания определенных параметров (для признаков в полуколичественных единицах).

Сравнение средних арифметических

Задача сравнения выборочных средних – это вопрос о том, действовал ли при составлении одной из выборок новый систематический фактор по

сравнению с другой выборкой. В терминах статистики отличия между средними могут иметь два противоположных источника:

1. Обе выборки взяты из одной генеральной совокупности, но средние отличаются в силу ошибки репрезентативности.

2. Выборки взяты из разных генеральных совокупностей, отличие средних вызвано в основном действием разных доминирующих факторов (а также и случайно).

Статистическая задача состоит в том, чтобы сделать обоснованный выбор. Исходно предполагается (Но): «Достоверных отличий между средними нет». Отличить закономерное от случайного можно только на основе знания законов поведения случайной величины. Для исключения чужеродных («выскакивающих») вариант мы применяли закон нормального распределения: в диапазоне четырех стандартных отклонений, $M \pm 1.96 \cdot S$, отклонение вариант от средней происходит по случайным причинам; за границами этого диапазона лежат чужеродные для данной выборки значения. Поскольку выборочные средние имеют нормальное распределение, критерий отличия двух выборочных средних также базируется на *свойствах нормального распределения*: в границах $M_{общ.} \pm 1.96 \cdot m$ (или приблизительно $M_{общ.} \pm 2 \cdot m$) выборочные средние арифметические отличаются от общей (генеральной) средней по случайным причинам. Тогда рабочая формула для t критерия отличия средних будет:

$$t = \frac{|M_1 - M_2|}{\sqrt{m_1^2 + m_2^2}} \sim t_{(\alpha,\, df)}.$$

Следует помнить, что разность средних нужно брать по модулю, т. е. без учета знака. Полученное этим способом значение критерия t Стьюдента сравнивают с табличным при выбранном уровне значимости (обычно для $\alpha = 0.05$) и числе степеней свободы (*объемы выборок без числа ограничений*, $df = n_1 + n_2 - 2$). Результатом такого сравнения должен стать один из двух вариантов следующего статистического вывода. Если полученное значение (величина) критерия больше табличного, значит, различия между параметрами при заданном уровне значимости и установленном числе степеней свободы достоверны. Если же полученная величина критерия меньше табличной, то при данном уровне значимости и числе степеней свободы различия между параметрами недостоверны. Последнее говорит о том, что различия случайны, никакого определенного вывода сделать нельзя, нулевая гипотеза остается неопровергнутой.

При сравнении выборочных параметров нормального и биномиального распределений используется одна и та же формула. Например, в процессе специальных исследований было установлено, что у стариков до лечения инсулином среднее содержание белков в крови составляло 81.04 ± 1.7, а после лечения – 79.33 ± 1.6. Нетрудно видеть, что полученные величины неодинаковы. Но достоверно ли это различие, закономерно ли оно? Можно ли на его основании утверждать, что лечение инсулином понижает содержание белков в крови? Ответ на этот вопрос может дать критерий достоверности различий

средних арифметических. Согласно общей нулевой гипотезе, средние не отличаются. Проверим ее с помощью критерия Стьюдента:

$$t = \frac{M_1 - M_2}{\sqrt{m_1^2 + m_2^2}} = \frac{81.04 - 79.33}{\sqrt{1.7^2 + 1.6^2}} = 0.7.$$

По таблице граничных значений критерия (табл. 6*П)* находим, что для уровня значимости $\alpha = 0.05$ и числа степеней свободы $df = 20 + 20 - 2 = 38$ величина критерия составляет $t_{(0.05,39)} = 2.03$. Поскольку полученное значение (0.7) меньше табличного (2.03), нулевая гипотеза сохраняется, различия между средними величинами статистически недостоверны (незначимы). Следовательно, влияние инсулина на содержание белков в крови приведенными выше данными не подтверждается и остается недоказанным, возможно, из-за недостаточного числа определений.

Сравнение долей

При сравнении достоверности различия долей или процентов (p) признаков, характеризующихся альтернативным распределением, применяют критерий Фишера с φ-преобразованием. Вместо процентов берут фи-значения $\varphi = \arcsin\sqrt{p}$ (или по таблице 10*П)* и подставляют их в формулу:

$$F = \frac{(\varphi_1 - \varphi_2)^2 \cdot n_1 \cdot n_2}{n_1 + n_2} \sim F_{(\alpha,\, df1,\, df2)},$$

где φ_1 и φ_2 – преобразованные доли,
n_1 и n_2 – объемы выборок.

Полученное значение сравнивают с табличным в соответствии с заданным уровнем значимости, $\alpha = 0.05$, и числом степеней свободы: $df_1 = 1$, $df_2 = n_1 + n_2 - 2$.

Например, в процессе учетов мелких млекопитающих в двух разных биотопах, где стояло по 200 ловушек, попалось соответственно 5 и 15 зверьков. Отличается ли численность животных на этих площадках? Если рассматривать ловушку как варианту, способную принимать два значения – «пустая» и «сработавшая» (со зверьком), то получаем выборку вариант (ловушек) с альтернативным распределением. Число пойманных особей можно пересчитать в процент сработавших ловушек: M_1 = 100% · 5 /200 = 2.5%, M_1 = 100% · 15 / 200 = 7.5%. По таблице 10*П* находим значения φ и вычисляем значение критерия: $F = \frac{(0.318 - 1.555)^2 \cdot 200 \cdot 200}{200 + 200} = 5.62$. Полученная величина (5.62) больше критической $F_{(0.05,\,1,\,398)} = 3.9$, значит, численность мелких млекопитающих во втором биотопе достоверно выше, чем в первом.

Сравнение показателей изменчивости

Наиболее точным методом определения достоверности различий между выборочными дисперсиями служит критерий F Фишера в форме отношения дисперсий (большее значение должно стоять в числителе):

$$F=\frac{S_1^2}{S_2^2}\sim F_{(\alpha,\,df1,\,df2)},$$

где $S_1 > S_2$, $df_1 = n_1 - 1$, $df_2 = n_2 - 1$.

Если полученная величина F больше табличного значения при принятом уровне значимости (табл. 7*П* для $\alpha = 0.05$ и табл. 8*П* для $\alpha = 0.01$) и числе степеней свободы (df_1 и df_2), то различие между дисперсиями признается достоверным; если она меньше, то расхождение между ними может считаться несущественным, случайным, т. е. нулевая гипотеза не отвергается.

Рассмотрим такой пример. При сравнении по показателю плодовитости (число эмбрионов на самку) двух популяций красной полевки с разным уровнем численности (у первой, горной, популяции плотность населения в два раза выше, чем у равнинной) оказалось, что при очень близких средних арифметических (соответственно $M_1 = 5.8$ и $M_2 = 5.4$, разница статистически недостоверна) стандартные отклонения значительно различаются: $S_1 = 1.82$, $S_2 = 0.52$ (при $n_1 = 27$, $n_2 = 12$). Отсюда

$$F=\frac{S_1^2}{S_2^2}=\frac{3.3124}{0.2704}=12.25\,.$$

Полученное значение критерия ($F = 12.2$) больше табличного $F_{(0.05,\,26,\,11)} = 2.6$, следовательно, нулевую гипотезу о случайности отличий можно отбросить, сделав вывод о том, что показатели изменчивости плодовитости в разных по численности популяциях достоверно отличаются. С биологических позиций это понятно, поскольку генетические отличия между особями практически по всем признакам, включая плодовитость, в больших популяциях выше, чем в малых. Новым фактором, усиливающим изменчивость особей в выборке, становится возможность появления аберрантных форм в условиях более свободной панмиксии.

Коэффициенты вариации также можно использовать для сравнения изменчивости разных показателей. Достоверность отличий коэффициентов оценивается с помощью критерия Стьюдента по формуле:

$$t=\frac{|CV_1-CV_2|}{\sqrt{m_1^2+m_2^2}}\sim t_{(0.05,\,n1+n2-2)},$$

где CV_1, CV_2 и m_1, m_2 – значения и ошибки коэффициентов вариации.

Вывод о достоверности отличий делается в том случае, если рассчитанное значение превысит табличное при заданном уровне значимости $\alpha = 0.05$ и числе степеней свободы $df = n_1 + n_2 - 2$. Сравним по критерию Стьюдента изменчивость веса тела землероек и плодовитости лисиц:

$CV_1 = 8.6 \pm 0.77\%$, $n_1 = 63$; $CV_2 = 26.7 \pm 2.2\%$, $n_2 = 76$, отсюда

$$t=\frac{|8.6-26.7|}{\sqrt{0.77^2+2.2^2}}=7.76.$$

Поскольку полученное значение (7.8) больше табличного ($t_{(0.05,\,137)} = 1.96$), изменчивость плодовитости лисиц достоверно выше, чем изменчивость веса тела землероек.

Сравнение выборок с помощью непараметрических критериев

Описанные выше статистические критерии (t, F и др.) относятся к *параметрическим*, т. к. *используют стандартные параметры распределений* (M, S, n). Они связаны с законом нормального распределения и применяются для оценки расхождения между генеральными параметрами по выборочным показателям сравниваемых совокупностей. Существенным достоинством параметрических критериев служит их большая статистическая мощность, т. е. широкие разрешающие возможности, а недостатком – трудоемкость расчетов, неприменимость к распределениям, сильно отклоняющимся от нормального, а также при исследовании качественных признаков.

Наряду с параметрическими критериями для ориентировочной оценки расхождений между выборками (особенно небольшими) применяются так называемые непараметрические критерии, ориентированные в первую очередь на исследование соотношений *рангов* исходных значений вариант. *Ранг – это число натурального ряда, которым обозначается порядковый номер каждого члена упорядоченной совокупности вариант.* Эта замена позволяет сравнивать выборки как по количественным, так и по качественным признакам, значения которых не имеют числового представления, но которые можно ранжировать. Конструкции непараметрических критериев отличаются простотой.

Вся процедура состоит из трех этапов – упорядочивание и ранжирование вариант, подсчет сумм рангов в соответствии с правилами данного критерия, сравнение полученной величины с табличным значением критерия. При этом с параметрическими критериями их роднит общая идеологическая подоплека. Нулевая гипотеза, как правило, состоит в том, что сравниваемые выборки взяты из одной и той же генеральной совокупности, значит, характер распределения вариант в этих выборках должен быть сходным. Поскольку вместо самих значений вариант используются ранги, все непараметрические методы исследуют один вопрос, насколько равномерно варианты разных выборок «перемешаны» между собой. Если варианты разных выборок более или менее регулярно чередуются в общем упорядоченном ряду, значит, они распределены сходным образом и отличий между совокупностями нет. Если же выборки пересекаются неполно (смешиваются только краями распределений, либо одна поглощает другую), то становится ясно, что эти выборки взяты из разных генеральных совокупностей (со смещенными центрами или разными дисперсиями).

Среди множества известных методов мы рассмотрим два метода: Уилкоксона – Манна – Уитни (довольно точный, но не самый простой для вычислений) и критерий Q Розенбаума. (простой для расчетов, но не очень точный).

Критерий U Уилкоксона – Манна – Уитни

Этот метод сравнения двух выборок признается наиболее чувствительным и мощным среди прочих непараметрических критериев. Согласно нулевой гипотезе, сравниваемые совокупности имеют одинаковые распределения.

Техника метода состоит в том, что все варианты сравниваемых совокупностей ранжируют в одном общем ряду: каждому значению присваивают ранг, порядковый номер. При этом одинаковым (повторяющимся) значениям вариант должен соответствовать один и тот же средний ранг (они как бы «делят места»). После этого ранги вариант суммируют отдельно по каждой выборке: $R_1 = \Sigma r_i$, $R_2 = \Sigma r_j$, $i = 1, 2, \ldots, n_1, j = 1, 2, \ldots, n_2$; $n = n_1 + n_2$
и вычисляют величину критерия:

$$t = \frac{U - 0.5 \cdot n_1 \cdot n_2}{\sqrt{(n_1 \cdot n_2 \cdot (n+1)/12)}},$$

где $U = \max(U_1, U_2)$ – максимальное значение из двух величин:

$$U_1 = n_1 \cdot n_2 + 0.5 \cdot n_1(n_1 + 1) - R_1,$$

$$U_2 = n_1 \cdot n_2 + 0.5 \cdot n_2(n_2 + 1) - R_2.$$

Если выборка достаточно велика ($n > 20$), величина статистики t сравнивается с табличным значением критерия Стьюдента для $df = \infty$ и $\alpha = 0.1$ (т. е. только для верхней 95%-й области нормального распределения). Считается, что метод хорошо работает для выборок объемом больше 10. В случае с меньшими выборками нужно пользоваться таблицами Уилкоксона – Манна – Уитни (табл. 11*П*).

В качестве примера сравним 5- и 35-дневных щенков песцов по активности фермента каталазы в сердце (E):
5-дневные: 41, 44, 31, 38, 43, 29, 71, 45, $M = 42.6$, $S = 12.8$, $n_1 = 8$,
35-дневные: 52, 51, 62, 52, 52, 50, 54, 62, 31, $M = 51.7$, $S = 9.0$, $n_2 = 9$.

Высокие коэффициенты вариации (30 и 17%) говорят о том, что распределения признаков, скорее всего, не соответствуют нормальному. Поэтому сравнивать средние следует с помощью непараметрического критерия. Ранжируем всю совокупность – упорядочим значения выборок по возрастанию:

E_5	29	31	38	41	43	44	45	71
E_{35}	31	50	51	52	52	52	54	62	62

Затем упорядочим все значения вместе, но так, чтобы значения каждой выборки располагались в двух отдельных рядах (E_5, E_{35}). Такое расположение упрощает назначение рангов (ряды r_5, r_{35}) и суммирование рангов (R):

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	R
E_5	29		31	38	41	43	44	45									71	
E_{35}		31							50	51	52	52	52	54	62	62		
r_5	1		2.5	4	5	6	7	8									17	50.5
r_{35}		2.5							9	10	12	12	12	14	15.5	15.5		102.5

$$U_1 = 9 \cdot 8 + 0.5 \cdot 9 \cdot (9+1) - 50.5 = 66.5,$$

$$U_2 = 9 \cdot 8 + 0.5 \cdot 8 \cdot (8+1) - 102.5 = 5.5,$$

$U = \max(U_1, U_2) = 66.5$, $n = 8 + 9 = 17$,

$$T = \frac{66.5 - 0.5 \cdot 9 \cdot 8}{\sqrt{(9 \cdot 8 \cdot 18/12)}} = 2.93.$$

Полученное значение (2.93) больше табличного ($t_{(0.1,\,\infty)} = 1.65$, табл. 6*П*), т. е. активность каталазы с возрастом меняется. Раз выборки малы, воспользуемся точными таблицами Уилкоксона – Манна – Уитни (табл. 11*П*). Получаем $t_{(0.05,\,n1,\,n2)} = t_{(0.05,\,8,\,9)} = 51$. Полученное значение (66.5) больше табличного (51), следовательно, различия между выборками достоверны.

Критерий Q Розенбаума

Этот критерий, как и предыдущие, оценивает достоверность различий двух эмпирических распределений, но в отличие от них почти не требует вычислений. Сравним два ряда цифр, характеризующих привесы (г) барашков одного возраста при добавлении в корм специальной подкормки (234, 277, 214, 201, 174, 167, 184, 157, 196, 173, 190, 191, 141, 150, 191) и без нее (183, 154, 175, 159, 157, 189, 198, 165, 176, 124, 173, 182, 204, 151, 147). Устанавливаем максимальные (277 и 204) и минимальные (141 и 124) значения и определяем порядковый номер сравниваемых совокупностей. В качестве *первой* следует принять выборку с *наибольшей вариантой* 277.

Далее находим число значений первой выборки, *превышающих* максимальное значение второй выборки (204): $Q_1 = 3$ (это варианты 234, 277, 214). Затем определяем число вариант второй выборки, *уступающих* по величине минимальному значению первой выборки (141): $Q_2 = 1$ (варианта 124). Далее определяем критерий Розенбаума как сумму полученных чисел: $Q = Q_1 + Q_2 = 3 + 1 = 4$. По таблице 12*П* находим критическое значение $Q_{(0.05,15,15)} = 6$. Поскольку эмпирическое значение (4) меньше табличного (6), приходим к выводу об отсутствии достоверного отличия выборок друг от друга, а значит, и влияния подкормки на привесы барашков. Следует все же иметь в виду, что возможности этого метода ограничены, он дает лишь прикидочный результат и оказывается эффективным только в случае сравнительно больших различий между выборками.

Сравнение двух частотных распределений. Критерий хи-квадрат

В практике биологических исследований часто бывает необходимо проверить ту или иную гипотезу, т. е. выяснить, насколько полученный экспериментатором фактический материал подтверждает теоретическое предположение, насколько анализируемые данные совпадают с теоретически ожидаемыми. Возникает задача статистической оценки разницы между фактическими данными и теоретическим ожиданием, установления того, в каких случаях и с какой степенью вероятности можно считать эту разницу достоверной и, наоборот, когда ее следует считать несущественной, незначимой, находящейся в пределах случайности. В последнем случае сохраняется гипотеза, на основе которой рассчитаны теоретически ожидаемые данные или показатели. Таким вариационно-статистическим приемом проверки гипотезы

служит метод *хи-квадрат* (χ^2). Этот показатель часто называют «критерием соответствия» или «критерием согласия» Пирсона. С его помощью можно с той или иной вероятностью судить о степени соответствия эмпирически полученных данных теоретически ожидаемым.

С формальных позиций сравниваются два вариационных ряда, две совокупности: одна – эмпирическое распределение, другая представляет собой выборку с теми же параметрами (n, M, S и др.), что и эмпирическая, но ее частотное распределение построено в точном соответствии с выбранным теоретическим законом (нормальным, Пуассона, биномиальным и др.), которому предположительно подчиняется поведение изучаемой случайной величины.

В общем виде формула критерия соответствия может быть записана следующим образом:

$$\chi^2 = \sum \frac{(a - A)^2}{A},$$

где a – фактическая частота наблюдений,
A – теоретически ожидаемая частота для данного класса.

Нулевая гипотеза предполагает, что достоверных различий между сравниваемыми распределениями нет. Для оценки существенности этих различий следует обратиться к специальной таблице критических значений хи-квадрат (табл. 9*П*) и, сравнив вычисленную величину χ^2 с табличной, решить, достоверно или не достоверно отклоняется эмпирическое распределение от теоретического. Тем самым гипотеза об отсутствии этих различий будет либо опровергнута, либо оставлена в силе. Если вычисленная величина χ^2 равна или превышает табличную $\chi^2_{(\alpha,\, df)}$, решают, что эмпирическое распределение от теоретического отличается достоверно. Тем самым гипотеза об отсутствии этих различий будет опровергнута. Если же $\chi^2 < \chi^2_{(\alpha,\, df)}$, нулевая гипотеза остается в силе. Обычно принято считать допустимым уровень значимости $\alpha = 0.05$, т. к. в этом случае остается только 5% шансов, что нулевая гипотеза правильна и, следовательно, есть достаточно оснований (95%), чтобы от нее отказаться.

Определенную проблему составляет правильное определение числа степеней свободы (df), для которых из таблицы берут значения критерия. Для определения числа степеней свободы из общего числа классов k нужно вычесть число ограничений (т. е. число параметров, использованных для расчета теоретических частот).

В зависимости от типа распределения изучаемого признака формула для расчета числа степеней свободы будет меняться. Для *альтернативного* распределения ($k = 2$) в расчетах участвует только один параметр (объем выборки), следовательно, число степеней свободы составляет $df = k - 1 = 2 - 1 = 1$. Для *полиномиального* распределения формула аналогична: $df = k - 1$. Для проверки соответствия вариационного ряда распределению *Пуассона* используются уже два параметра – объем выборки и среднее значение (численно совпадающее с дисперсией); число степеней свободы $df = k - 2$. При проверке

соответствия эмпирического распределения вариант *нормальному* или *биномиальному* закону число степеней свободы берется как число фактических классов минус три условия построения рядов – объем выборки, средняя и дисперсия, $df = k-3$. Сразу стоит отметить, что критерий χ^2 работает только для выборок *объемом не менее 25 вариант*, а частоты отдельных классов должны быть *не ниже 4*.

Вначале проиллюстрируем применение критерия хи-квадрат на примере анализа *альтернативной изменчивости*. В одном из опытов по изучению наследственности у томатов было обнаружено 3629 красных и 1176 желтых плодов. Теоретическое соотношение частот при расщеплении признаков во втором гибридном поколении должно быть 3:1 (75% к 25%). Выполняется ли оно? Иными словами, взята ли данная выборка из той генеральной совокупности, в которой соотношение частот 3:1 или 0.75:0.25?

Сформируем таблицу (табл. 4), заполнив значениями эмпирических частот и результатами расчета теоретических частот по формуле:

$A = n \cdot p,$

где p – теоретические частости (доли вариант данного типа),

n – объем выборки.

Например, $A_2 = n \cdot p_2 = 4805 \cdot 0.25 = 1201.25 \approx 1201$.

Таблица 4

Значение (цвет плода), x_j	Фактическая частота, a	Теоретическая частость, p	Теоретическая частота, A	$\frac{(a-A)^2}{A}$
Красный	3629	0.75	3604	0.187621
Желтый	1176	0.25	1201	0.5204
Сумма	$n = \Sigma a = 4805$	1	$n = \Sigma A = 4805$	$\chi^2 = 0.71$

Далее вычисляем хи-квадрат $\chi^2 = 0.71$ и число степеней свободы (при двух классах и одном ограничении, объеме выборки) $df = k - 1 = 2 - 1 = 1$. По табл. 9*П* находим критическое значение $\chi^2_{(0.05,\ 1)} = 3.84$. Поскольку полученная величина (0.71) меньше табличной (3.84), различия сравниваемых распределений статистически недостоверны. Иначе говоря, фактические частоты хорошо согласуются с теоретически ожидаемыми. Полученные данные не отвергают принятую гипотезу о том, что в нашем случае имеется соотношение 3:1.

Здесь следует еще раз обратить внимание читателей на то обстоятельство, что сохранение нулевой гипотезы нельзя считать доказательством справедливости нулевой гипотезы. Результатами представленных вычислений теория о расщеплении по фенотипам в соотношении 3:1 *не доказана*, хотя и не опровергнута. Статистика доказывает только факт отличий, но не их отсутствие. Чтобы доказать теорию, нужно предположить антитеорию (например, соотношение 1:1) и опровергнуть ее с помощью статистических приемов.

В процессе другого исследования добыты 671 самец и 569 самок. Требуется определить, подтверждают ли эти данные факт преобладания самцов или налицо просто случайное отличие цифр. Теоретическое отношение признаков (соотношение полов) 1:1. Подтверждается ли оно? Находим сумму 671+569=1240, среднее 620, $\chi^2 = \frac{(671-620)^2}{620} + \frac{(569-620)^2}{620} = 8.4.$

Сравнение вычисленного (8.4) и критического значений (для $df = 1$ и $\alpha = 0.05$ $\chi^2_{(0.05,\,1)}$ = 3.84) явно свидетельствует о существенном отклонении фактического соотношения полов от гипотезы – 1:1. Вероятность правильности нулевой гипотезы (т. е. что в данном случае действительно имеет место численное равенство полов) оказалась даже меньше 0.01. Следовательно, есть все основания говорить о достоверном преобладании самцов.

В качестве первого примера решим задачу, соответствует ли *закону Пуассона* распределение числа повторных отловов альбатросов (табл. 5). В этом случае рассматривается процесс, этапами которого выступают события «отлов птицы». В чреде таких событий встречаются редкие – «отлов меченной особи». Биологическая подоплека состоит в следующем: случайны ли повторные отловы птиц или есть факторы, ответственные за нарушение случайности? Например, птицы могут приманиваться и стремиться попасться вновь либо могут стараться избежать повторного отлова. В обоих случаях птицы будут «умышленно» попадаться чаще или реже, нарушая *случайность* повторного отлова и искажая тем самым форму распределения, которое будет отходить от формы, предписанной законом Пуассона. Согласно нулевой гипотезе, птицы ведут себя случайно, их встречаемость соответствует этому закону. Алгоритм расчетов теоретических частот для этого случая прост и основан на формулах прямого расчета теоретических частот:

$$A_0 = \frac{n}{e^M} \text{ (частота нулевого класса),}$$

$$A_x = \frac{M}{x} \cdot A_{x-1} \text{ (частота прочих классов),}$$

где M – средняя арифметическая ряда,
x – значение ряда (число объектов в пробе),
A_x – теоретическая частота значения x,
n – объем выборки (число проб),
e = 2.7183… – основание натурального логарифма.

Параметры данного вариационного ряда были рассчитаны выше (с. 23): M = 0.968. Теоретическая частота нулевого значения равна:

$$A_0 = \frac{n}{e^M} = \frac{32}{e^{0.968}} = 11.93803 \approx 12,$$

частота значения $x = 1$:

$$A_x = \frac{M}{x} \cdot A_{x-1} = \frac{0.968}{1} \cdot 11.93 = 11.55602 \approx 11$$

и т. д. (табл. 5, графа 3).

Таблица 5

Число повторных отловов, *x*	Фактическая частота, *a*	Теоретическая частота, *A*	$\frac{(a-A)^2}{A}$
0	15	12	0.75
1	7	11	1.45
2	7	6	
3	2 } 10	2 } 9	0.17
4	1	1	
Сумма	$n = \Sigma a = 32$	$n = \Sigma A = 32$	$\chi^2 = 2.31$

По окончании вычислений получаем два ряда частот, отличия между которыми оцениваются по критерию хи-квадрат.

Перед расчетом значения критерия следует убедиться, что выполнены требования к данным для расчета критерия χ^2:

– объем выборки более 25 вариант, $n > 25$,

– суммы эмпирических и теоретических частот равны объему выборки $n = \Sigma a = \Sigma A$ (с точностью не ниже 1–2%),

– все классы эмпирического и теоретического рядов имеют частоты более 4, $a_j > 4$; если какие-либо классы имеют меньше 4 вариант (у нас значения 3 и 4 имеют частоты 2 и 1), то они должны быть объединены (суммированы) с соседними, что и показано в таблице с помощью фигурных скобок.

Далее вычисляем значения критерия: для первой строки

$$\frac{(a-A)^2}{A} = \frac{(15-12)^2}{12} = 0.75$$

и т. д. (графа 4), итого $\chi^2 = 2.31$.

Число степеней свободы находим как *число окончательных классов* (3) минус число ограничений: $df = k - 2 = 3 - 2 = 1$.

Табличное значение $\chi^2_{(0.05,1)} = 3.84$. Полученная величина (2.31) меньше табличной (3.84), следовательно, нулевая гипотеза не отвергается: эмпирическое распределение достоверно не отличается от распределения Пуассона. Иными словами, у нас нет оснований утверждать, что вероятность повторного отлова изменяется: нельзя утверждать, что сама операция отлова привлекает или пугает птиц.

Соответствие эмпирического ряда *распределению Пуассона* можно проверить и другим способом: сравнив по критерию Фишера величины средней арифметической и дисперсии для числа степеней свободы $df_1 = n - 1$, $df_2 = n - 1$. В нашем случае $M = 0.968$, $S^2 = 1.257$, $F = 1.257 / 0.968 = 1.157$. Поскольку эта величина меньше табличной ($F_{(0.05,\,31,\,31)} = 1.84$), сравниваемые показатели достоверно не отличаются, а равенство средней и дисперсии характерно лишь для распределения Пуассона.

При статистическом исследовании непрерывных признаков нужно быть уверенным, что они действительно подчиняются *нормальному закону*, а

в случае дискретных признаков – биномиальному. Для такой проверки нулевая гипотеза звучит так: «полученное распределение соответствует нормальному (биномиальному)» или «выборка взята из генеральной совокупности, подчиняющейся закону нормального (биномиального) распределения». Все вычислительные операции для случаев нормального и биномиального распределений совпадают. Рассмотрим проверку нулевой гипотезы: распределение землероек по массе тела (см. пример на стр. 13) подчиняется нормальному закону.

Расчеты начинаются с построения вариационного ряда и поиска центральных значений для каждого класса (табл. 6 и 7). Далее по формуле $t = \frac{|x_j - M|}{S}$ вычисляются нормированные отклонения середины каждого классового интервала (x_j) от общей средней M (S – стандартное отклонение). В нашем случае $M = 9.29$ г, $S = 0.897$ г., тогда, например, для второго интервала получаем: $t = |8.05 - 9.27| / 0.897 = 1.38$. Далее определяем теоретические частости нормального распределения, или ординаты нормальной кривой (табл. 4*П*), соответствующие вычисленным нормированным отклонениям. Для $t = 1.38$ находим $p = 0.1539 \approx 0.15$ (табл. 6, графа 5). (Заметим, что модуль в формуле нормированных отклонений берется потому, что в таблице 6*П* приведены частости p только для положительных значений t.) Следующая операция, вычисление теоретических частот, ведется по формуле:

$A = c \cdot p$,

где p – ординаты нормальной кривой,

c – константа ряда, определяемая по формуле $c = \frac{dx \cdot n}{S}$,

dx – классовый интервал (в данном случае он равен 0.7) (см. с. 13),

n – объем выборки (63).

Для нашего примера $c = \frac{0.7 \cdot 63}{0.897} = 49.16$.

Таблица 6

Классовые интервалы	Центр интервала, x_j	Фактическая частота, a	Нормированное отклонение, t	Ординаты нормальной кривой, p	Теоретическая частота, A	$\frac{(a-A)^2}{A}$
7 – 7.7	7.35	2	2.16	0.04	2	
7.8 – 8.4	8.05	7 } 9	1.38	0.15	8 } 10	0.1
8.5 – 9.1	8.75	18	0.60	0.33	17	0.25
9.2 – 9.8	9.45	22	0.18	0.39	19	0.47
9.9 – 10.5	10.15	10	0.96	0.25	12	0.33
10.6 – 11.2	10.85	1	1.74	0.09	4	
11.3 – 11.9	11.55	3 } 4	2.52	0.02	1 } 5	0.2
Сумма		$n=\Sigma a= 63$			$n=\Sigma A= 63$	$\chi^2 = 1.36$

Теоретическая частота для $f = 0.15$ составит:

$A = 49.16 \cdot 0.1539 = 7.55 \approx 8$ (графа 6).

В результате вычислений получаем теоретическое нормальное распределение с параметрами $M = 9.29$ г, $S = 0.897$ г, $n = 63$ (см. рис. 4, с. 20).

Перед расчетом критерия хи-квадрат проверяем совпадение суммы эмпирических и теоретических частот (по 63 варианты) и минимальные объемы в отдельных классах. Поскольку в крайних классах частоты были ниже 4, проводим их объединение (отмечено скобками), после чего число классов сократилось до $k = 5$. Вычисляем значения χ^2: для первого класса $(9 - 10)^2 / 10 = 0.1$, для всего ряда $\chi^2 = 1.36$. Число степеней свободы $df = 5 - 3 = 2$. Табличное значение (табл. 9*П*) $\chi^2_{(0.05,\,2)} = 5.99$.

Поскольку полученное значение (1.36) меньше табличного (5.99), нулевая гипотеза сохраняется, распределение бурозубок по массе тела достоверно от нормального не отличается.

Аналогичные расчеты для дискретного признака (плодовитость лисиц), имеющего предположительно *биномиальное распределение* (дискретный аналог нормального), представлены в табл. 7. Так, при параметрах $M = 5$ экз., $S = 1.33$ экз. для второго интервала получаем: $t = |8 - 5| / 1.33 = 1.5$.

Таблица 7

Центр интервала, x_j	Фактическая частота, a	Нормированное отклонение, t	Ординаты нормальной кривой, p	Теоретическая частота, A	$\frac{(a-A)^2}{A}$
2	1 } 9	2.26	0.031	2 } 9	0
3	8	1.5	0.129	7	
4	16	0.75	0.301	17	0.05
5	23	0	0.399	23	0
6	21	0.75	0.301	17	0.94
7	3 } 6	1.5	0.129	7 } 9	1
8	3	2.26	0.031	2	
Сумма	$n = \Sigma a = 75$			$n = \Sigma A = 75$	$\chi^2 = 2$

Соответствующая ордината нормальной кривой равна $p = 0.1295$ (графа 4), теоретическая частота составит:

$A = c \cdot p = 56.38 \cdot 0.129 = 7.3 \approx 7$ (графа 5),

поскольку значение $c = 1 \cdot 75 / 1.33 = 56.38$. В результате вычислений получаем частоты (A) распределения (с параметрами $M = 5$, $S = 1.33$, $n = 75$), строго соответствующего биномиальному (см. рис. 5, с. 21). Объединим классы с частотами менее 4 и рассчитаем значение критерия $\chi^2 = 2$. Число степеней свободы (при трех ограничениях и пяти классах) равно $df = 5 - 3 = 2$. Поскольку это значение ($\chi^2 = 2$) меньше табличного ($\chi^2_{(0.05,\,2)} = 5.99$), нулевая гипотеза не может быть отклонена, значит, распределение лисиц по плодовитости в целом соответствует биномиальному закону.

ОЦЕНКА ВЛИЯНИЯ ФАКТОРА

При изучении и анализе сложных и многообразных причинно-следственных отношений между объектами и явлениями биологу приходится учитывать целый комплекс внешних и внутренних факторов, от которых в конечном итоге зависят уровень и ход наблюдаемых процессов, те или иные биологические свойства живых организмов, их динамика и разнообразие. При этом зачастую важно оценивать не только роль одного из многочисленных внешних факторов, но и их взаимодействие при констелляционном влиянии на популяцию или организм.

Идейная база для изучения действия факторов содержится уже в методе сравнения двух выборок. Биологическим содержанием операции сравнения двух выборок, в конце концов, выступает поиск факторов, ответственных за смещение средних арифметических или усиление изменчивости признаков. Развивая это направление биометрического исследования, можно не ограничиваться только двумя «дозами» фактора, но изучить серию ситуаций, в которых фактор проявлял разную силу действия на результативный признак – от самого слабого до самого сильного. При этом каждому уровню фактора будет соответствовать отдельная выборка и общая задача получит формулировку «сравнить несколько выборок». В терминах факториальной биометрии вопрос о влиянии фактора на признак звучит так: сказывается ли отличие условий получения разных выборок на качестве (значениях) вариант? В терминах статистики вопрос звучит несколько иначе: из одной ли генеральной совокупности отобраны все выборки, оценивают ли выборочные средние арифметические одну и ту же генеральную среднюю? Вариантов ответа может быть только два:

1. Все выборки отобраны из одной генеральной совокупности, условия возникновения вариант одни и те же.
2. Выборки отобраны из разных генеральных совокупностей, условия возникновения вариант выборок различаются.

В постановке вопроса можно уловить противоречие. Выше было сказано, что по условию задачи выборки формировались в разных условиях, и тут же предполагается, что условия были одинаковые. На самом деле противоречия нет, поскольку речь идет об определении чувствительности признака к действию фактора. Условия формирования выборок могут отличаться, но они могут никак и не сказаться на величине изучаемого признака, не отразиться на значениях вариант. Смысл статистического сравнения в том и состоит, чтобы оценить эффективность действия фактора на признак, доказать реальность реакции вариант выборок на разные условия их формирования. В сферу исследования можно вовлекать как один, так и два признака, как количественные, так и качественные характеристики. В каждом случае процедура анализа несколько отличается.

Однофакторный дисперсионный анализ количественных признаков

Дисперсионный анализ позволяет оценить степень и достоверность отличия нескольких выборочных средних одновременно, т. е. изучить влияние одного контролируемого фактора на результативный признак путем оценки его относительной роли в общей изменчивости этого признака, вызванной влиянием всех факторов. Сущность дисперсионного анализа заключается в расчленении общей вариации (дисперсии) изучаемого признака, вычисляемой по сумме квадратов отклонений отдельных вариант (x) от средней арифметической всего комплекса наблюдений (M), на его составные части – дисперсию, вызванную организованными, учитываемыми в исследовании факторами (факториальную дисперсию), оценивающую межгрупповую изменчивость, и дисперсию, обусловленную остальными, неорганизованными в данном исследовании факторами (внутригрупповую, или случайную, дисперсию) отклонения отдельных значений от средней в группе.

Общая вариация (сумма квадратов) признака рассчитывается как сумма квадратов отклонений всех вариант (x_i) от общей средней (M):

$$C_{общ.} = \Sigma\ (x_i - M)^2.$$

Факториальная (межгрупповая, межвыборочная) сумма квадратов рассчитывается как сумма квадратов отклонений частных средних (M_i) для каждой выборки (всего k выборок) от общей средней:

$$C_{факт.} = \Sigma\ (M_j - M)^2.$$

Остаточная (случайная, внутригрупповая) сумма квадратов есть сумма квадратов отклонений вариант каждой выборки (x_i) от своей средней (M_j):

$$C_{случ.} = \Sigma\ (x_i - M_j)^2.$$

Очевидно, что в общем комплексе наблюдений должно выполняться равенство $C_{общ.} = C_{факт.} + C_{случ.}$

Отношение сумм квадратов к соответствующему числу степеней свободы дает оценку величины дисперсии, или средний квадрат, иногда ее именуют варианса. Влияние изучаемого фактора отражает факториальная, или межгрупповая, дисперсия $S^2_{факт.}$, а влияние случайных неорганизованных в данном исследовании причин – случайная $S^2_{случ.}$, или внутригрупповая, остаточная дисперсия $S^2_{остат.}$:

$$S^2_{факт.} = \sum^{k} (M_j - M_{общ.})^2 / df_{факт.},$$

где $df_{факт.} = k - 1, j = 1, 2, \ldots, k$, k – число сравниваемых средних.

$$S^2_{случ.} = \sum^{k} \sum^{n_j} (x_{ij} - M_j)^2 / df_{случ.},$$

где $df_{случ.} = n - 1, i = 1, 2, \ldots, n$, n – число вариант всех выборок.

Сила влияния фактора определяется как доля частной суммы квадратов в общем варьировании признака. Показатель силы влияния изучаемого фактора составляет: $\eta^2_{факт.} = C_{факт.} / C_{общ.}$, неорганизованных (случайных): $\eta^2_{случ.} = C_{случ.} / C_{общ.}$; сумма этих показателей, естественно, равна единице: $\eta^2_{факт.} + \eta^2_{случ.} = 1$. Заметим, что показатель силы влияния дисперсионного

комплекса есть не что иное, как квадрат пирсоновского корреляционного отношения, которым и оценивается относительная доля влияния организованного (изучаемого) фактора в общем суммарном статистическом влиянии всех факторов, определяющих развитие данного результативного признака.

О достоверности оценок влияния факторов судят по уже знакомому нам критерию Фишера: $F = \frac{S^2_{факт.}}{S^2_{случ.}} \sim F_{(\alpha,\, df1,\, df2)}$,

где $df_1 = k - 1$, $df_2 = n - k$, k – число градаций,
n – общий объем всех выборок.

Проверяется нулевая гипотеза: «влияние фактора на признак отсутствует». Влияние считается доказанным, если величина расчетного критерия равна или превышает свое табличное значение с принятым уровнем значимости (обычно $\alpha = 0.05$) (F определяется по табл. *7П*). Все параметры однофакторного дисперсионного анализа и порядок их вычислений представлены в таблице 8.

Таблица 8

Составляющие дисперсии	Суммы квадратов (*SS*), *C*	Сила влияния, η^2	Степени свободы, *df*	Дисперсии (средний квадрат, *MS*), S^2	Критерий влияния, *F*
Факториальная	$C_{факт.} = \Sigma (M_j - M)^2$	$\frac{C_{факт.}}{C_{общ.}}$	$k - 1$	$S^2_{факт.} = \frac{C_{факт.}}{df_{факт.}}$	$F = \frac{S^2_{факт.}}{S^2_{случ.}}$
Случайная	$C_{случ.} = \Sigma (x_i - M_j)^2$		$n - k$	$S^2_{случ.} = \frac{C_{случ.}}{df_{случ.}}$	
Общая	$C_{общ.} = \Sigma (x_i - M)^2$				

Однофакторным называется анализ, изучающий действие на результативный признак только одного организованного фактора *А*. Для примера оценим влияние растворенного в воде вещества на плодовитость дафний, используемых в качестве тест-объектов в водно-токсикологических экспериментах. В ходе предварительного исследования были получены четыре выборки, четыре группы значений плодовитости животных, выращенных в средах с разным содержанием химической добавки.

Сначала необходимо сгруппировать выборочный материал в комбинативную таблицу (организовать дисперсионный комплекс). Для этого варианты каждой выборки записываются в отдельные графы, именуемые градациями (табл. 9). Результативным признаком служит средняя плодовитость дафний за неделю (для иллюстративности расчетов она дана в целых числах). В

нашем примере организованы 4 градации – чистая вода (контроль, градация *A*1; значения плодовитости 6, 5, 5, 7), слабая концентрация вещества (5 мг/л, *A*2; 8, 7, 6, 6), средняя (15 мг/л, *A*3; 8, 8, 7) и сильная (30 мг/л, *A*4; 8, 7, 9). Предлагаемый ниже алгоритм расчетов позволяет использовать неравное число вариант в градациях. Расчеты показаны в таблице 9.

Таблица 9

	Градации фактора									
	*A*1		*A*2		*A*3		*A*4			
	x	x^2	x	x^2	x	x^2	x	x^2		
	6	36	8	64	8	64	8	64		
	5	25	7	49	8	64	7	49		
	5	25	6	36	7	49	9	81		
	7	49	6	36						
									Σ	
Σx^2		135		185		177		194	691	$H1 = \Sigma\Sigma x^2 = 691$
Σx	23		27		23		24		97	$H2 = (\Sigma\Sigma x)^2/n =$
n	4		4		3		3		14	$= (97)^2/14 = 672$
$\Sigma x^2/n$	132		182		176.3		192		682.8	$H3 = \Sigma\Sigma x^2/n =$ $= 682.8$
M	5.8		6.8		7.67		8		6.93	

$$C_{факт.} = H3 - H2 = 682.8 - 672 = 10.76$$
$$C_{случ.} = H1 - H2 = 691 - 672 = 8.17$$
$$C_{общ.} = H1 - H3 = 691 - 682.8 = 18.93$$

Полученные значения позволяют вычислить дисперсии, определить силу влияния фактора и критерий достоверности Фишера.

Составляющие дисперсии	Суммы квадратов, C	Сила влияния, η^2	Степени свободы, df	Дисперсии, S	Критерий, F
Факториальная	10.76	57%	3	3.59	4.39
Случайная	8.17		10	0.82	
Общая	18.93			4.39	

Поскольку полученное значение критерия ($F = 4.39$) больше табличного ($F_{(0.05,3,10)} = 3.7$) (табл. 7*П*), отличие факториальной и случайной дисперсий достоверно, влияние фактора значимо.

Отсюда следует биологический вывод: стимулирующее влияние изучаемого фактора (вещества) на плодовитость дафний относительно велико (57%) и достоверно (с вероятностью $P > 0.95$).

Непараметрический однофакторный дисперсионный анализ

Рассмотренные выше схемы дисперсионного анализа исходили из предположения о нормальном распределении изучаемого результативного признака. Когда для какого-либо признака нет уверенности, что выполняется предположение о его нормальном распределении, когда требуется провести анализ быстро и без особой точности, когда мало данных или они выражены *качественными признаками*, можно использовать схему непараметрического дисперсионного анализа. Этот метод более неприхотлив, но менее точен, нежели параметрический анализ. Он исследует распределения вариант в нескольких выборках. Нулевая гипотеза состоит в том, что распределения одинаковы, т. е. выборки взяты из одной генеральной совокупности.

Порядок вычислений состоит в том, что все варианты ранжируются в порядке возрастания. Затем суммируются ранги вариант по каждой выборке отдельно и рассчитывается критерий:

$$H = \frac{12}{n\cdot(n-1)}\cdot\left(\frac{R_1^{\;2}}{n_1}+\ldots+\frac{R_j^{\;2}}{n_j}+\ldots+\frac{R_k^{\;2}}{n_k}\right)-3\cdot(n+1)\ \sim\chi^2{}_{(\alpha,\,k-1)},$$

где n – число всех вариант,
n_j – объем j-й градации фактора,
R_j – сумма рангов для каждой j-й градации фактора,
k – число градаций фактора ($j = 1, 2, \ldots, k$).

При объеме выборок больше 5 вариант статистика H имеет распределение хи-квадрат с $df = k - 1$ степенями свободы и сравнивается со значениями из табл. 9*П*.

Применим эту схему (табл. 10) к нашим данным из табл. 9, расположив их в строку.

№ п/п	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Градация	1	1	1	1	2	2	2	2	3	3	3	4	4	4
Значение	5	5	6	7	6	6	7	8	7	8	8	7	8	9

Затем упорядочим и ранжируем их. Для нескольких одинаковых значений берется средний ранг.

№ п/п	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Градация	1	1	1	2	2	1	2	3	4	2	3	3	4	4
Значение	5	5	6	6	6	7	7	7	7	8	8	8	8	9
Ранг	1.5	1.5	4	4	4	7.5	7.5	7.5	7.5	11.5	11.5	11.5	11.5	14

Наконец, разнесем ранги по градациям и подсчитаем необходимые суммы.

Таблица 10

Градация	1	1	1	1	2	2	2	2	3	3	3	4	4	4
Значение	5	5	6	7	6	6	7	8	7	8	8	7	8	9
Ранг, R	1.5	1.5	4	7.5	4	4	7.5	11.5	7.5	11.5	11.5	7.5	11.5	14
Сумма, R				14.5				27			30.5			33
n				4				4			3			3
R^2/n				52.56				182.3			310.1			363

Общий объем выборки равен n = 14. Величина критерия H составит:

$$H = \frac{12}{14 \cdot 13} \cdot (52.56 + 182.3 + 310.1 + 363) - 3 \cdot 13 =$$

$$= 0.065934 \cdot 907.8958 - 45 = 14.86.$$

По таблице распределения статистики χ^2 для $\alpha = 0.05$ и $df = 4 - 1 = 3$ находим $\chi^2_{(0.05,\ 3)} = 7.81$. Полученное значение критерия (14.86) больше табличного (7.81), значит, отличие выборочных распределений достоверно. Химическая добавка действительно изменяет плодовитость дафний.

Двухфакторный дисперсионный анализ количественных признаков

Двухфакторный дисперсионный анализ исследует влияние на результативный признак двух факторов как порознь, так и совместно. Учет эффекта влияния каждого фактора по отдельности теоретически ничем не отличается от описанных выше схем. И там и тут оценивается изменчивость средних по градациям на фоне случайной изменчивости вариант внутри градаций, с помощью критерия Фишера устанавливается достоверность отличий межгрупповых дисперсий от внутригрупповых.

Двухфакторный дисперсионный анализ, естественно, требует более сложных вычислительных операций, чем однофакторный, но в принципе ничем не отличается от описанных выше схем. Однако это относится лишь к ортогональным (равномерным, или пропорциональным) комплексам, характеризующимся равной или, по крайней мере, пропорциональной численностью групп (в градациях содержатся одинаковые или пропорциональные числа вариант). Что же касается неортогональных многофакторных комплексов, то их анализ принципиально возможен, но имеет свои особенности, существенно усложняющие технику вычислений, и в настоящем пособии не рассматривается.

На практике вполне допустим и такой способ избегнуть сложностей обработки неравномерных комплексов, как искусственное превращение их в равномерные. Для этого нужно составить выборки одинаковой или пропорциональной численности, используя только часть имеющихся данных. Следует, однако, помнить, что такой отбор не должен быть субъективным. Чтобы не допустить возможной тенденциозности, лучше всего прибегнуть к жеребьевке.

Важным преимуществом двухфакторного дисперсионного анализа перед однофакторным служит то, что с его помощью удается определить варьирование по сочетанию градаций $C_{сочет.} = C_{AB}$, позволяющее получить новый

и весьма ценный в биологическом отношении показатель – оценку влияния сочетанного действия (взаимодействия) факторов.

Общая вариация (сумма квадратов) признака теперь состоит из четырех компонентов за счет более детального разложения факториальной дисперсии.

Правило разложения вариаций предстает как:

$C_{общ.} = C_A + C_B + C_{AB} + C_{случ.}$,

$C_{факт.} = C_{общ.} - C_{случ.} = C_A + C_B + C_{AB}$.

Для расчетов используются следующие смысловые формулы:

$C_{общ.} = \Sigma(x_i - M)^2$,

$C_{A.} = \Sigma(M_{Aj} - M)^2$, j – число градаций фактора A, M_{Aj} – групповые средние по градациям фактора А,

$C_B = \Sigma(M_{Bk} - M)^2$, k – число градаций фактора B, M_{Bk} – групповые средние по градациям фактора В,

$C_{случ.} = \Sigma(x_i - M_{xi})^2$,

$C_{AB} = C_{общ.} - (C_A + C_B + C_{случ.})$.

Сочетанное действие (взаимодействие) каждого из двух факторов проявляется в усилении или ослаблении непосредственного действия другого фактора на объект исследования. К примеру, неурожай кормов усугубляет негативное действие зимнего холода на численность популяций мелких млекопитающих.

Рассмотрим числовой пример – испытания стимулятора многоплодия при разной полноценности рационов. Полноценность рациона (первый фактор) представлена двумя градациями: $A1$ – рацион с недостатком минеральных веществ, $A2$ – рацион, полностью сбалансированный по всем питательным веществам, включая и минеральные. Стимулятор (второй фактор) был испытан в трех дозах: $B1$ – одинарная, $B2$ – двойная, $B3$ – тройная. Результативный признак – плодовитость самок, измерявшаяся числом детенышей в помете. Для каждого сочетания градаций рациона и стимулятора были подобраны три одновозрастные самки.

Комбинативная таблица двухфакторного равномерного дисперсионного комплекса с трехкратной повторностью ($n_i = 3$) включает две градации по фактору A и три градации по фактору B (табл. 11). Варианты размещаются по градациям, определяется объем градации, вычисляются суммы вариант, частные средние, затем вспомогательные величины ($H1$, $H2$, $H3$, H_A, H_B) и суммы квадратов отклонений (дисперсий) по рабочим формулам. В завершение всего заполняют таблицу дисперсионного анализа (табл. 12), находят показатель достоверности влияния Фишера и, сопоставляя его с табличным для соответствующих степеней свободы и принятого уровня значимости, делают статистический вывод.

Таблица 11

Градации факторов		$A1$		$A2$			Для B		
		x	x^2	x	x^2	Σ	M_B	$\Sigma\Sigma x^2/n$	$\Sigma(\Sigma x^2/n)$
$B1$		5	25	1	1				
		6	36	4	16				
		7	49	1	1				
	Σx^2		110		18	$\Sigma\Sigma x^2 = 128$			
	Σx	18		6		$\Sigma\Sigma x = 24$	4	96	
	n	3		3		$n_{B1} = 6$			
	$\Sigma x^2/n$	108		12		$\Sigma(\Sigma x^2/n) = 120$			
$B2$		4	16	10	100				
		3	9	9	81				
		5	25	11	121				$\boldsymbol{H_B} =$
	Σx^2		50		302	$\Sigma\Sigma x^2 = 352$			$\Sigma(\Sigma x^2/n)$
	Σx	12		30		$\Sigma\Sigma x = 42$	7	294	$= 486$
	n	3		3		$n_{B2} = 6$			
	$\Sigma x^2/n$	48		300		$\Sigma(\Sigma x^2/n) = 348$			
$B3$		2	4	7	49				
		3	9	4	16				
		1	1	7	49				
	Σx^2		14		114	$\Sigma\Sigma x^2 = 128$			
	Σx	6		18		$\Sigma\Sigma x = 24$	4	96	
	n	3		3		$n_{B3} = 6$			
	$\Sigma x^2/n$	12		108		$\Sigma(\Sigma x^2/n) = 120$			
$\Sigma\Sigma$	$\Sigma\Sigma x^2$		174		434	$\boldsymbol{H1} = \Sigma\Sigma\Sigma x^2 = 608$			
	$\Sigma\Sigma x$	36		54		$\Sigma\Sigma\Sigma x = 90$	$\boldsymbol{H2} =$	$(\Sigma\Sigma\Sigma x)^2/\text{N} = 450$	
	$n_A = \Sigma n$	9		9		$N = \Sigma\Sigma n = 18$			
	$\Sigma x^2/n$	168		420		$\boldsymbol{H3} = \Sigma\Sigma(\Sigma x^2/n) = 588$			
Для А	$M_A = \Sigma\Sigma x/n$	2		6		$j = 2$ – число градаций фактора А			
	$\Sigma x^2/n$	144		324		$k = 3$ – число градаций фактора В			
	$\boldsymbol{H_A} = \Sigma(\Sigma x^2/n) = 468$								

$C_{общ.} = H1 - H2 = 608 - 450 = 158$
$C_{случ.} = H1 - H3 = 608 - 588 = 20$
$C_{факт.} = C_{A+B+AB} = H3 - H2 = 588 - 450 = 138$
$C_A = H_{\text{А}} - H2 = 468 - 450 = 18$
$C_B = H_B - H2 = 486 - 450 = 36$
$C_{AB} = C_{факт.} - C_A - C_B = 138 - 18 - 36 = 84$

В нашем примере все факториальные влияния оказались достоверными с доверительной вероятностью $P > 0.95$ (табл. 12). Это позволяет сделать определенные выводы относительно действия стимулятора на плодовитость самок. Влияние каждого фактора в отдельности (качества рациона и дозы стимулятора) и их суммарного эффекта достаточно существенно, но особенно результативно действие стимулятора в сочетании с полноценным рационом (величина η^2_{AB} выше, чем η^2_A и η^2_B). Более того, при недостатке в корме

минеральных веществ двукратные и трехкратные дозы стимулятора могут даже снизить плодовитость животных.

Таблица 12

Составляющие дисперсии	Суммы квадратов, С	Сила влияния, η^2 (%)	Степени свободы, *df*	Дисперсии, *S*	Критерий, *F* ($F_{(\alpha, dfi, dfсл.)}$)
Фактор *A*	18	11	$j-1=1$	18	10.8 (4.7)
Фактор *B*	36	23	$k-1=2$	18	10.8 (3.9)
Взаимодействие *AB*	84	53	$df_A \cdot df_B = 2$	42	25.2 (3.9)
Факториальная (всего)	138	87	$j \cdot k-1=5$	27.6	16.5 (3.1)
Случайная	20	13	$N-j \cdot k=12$	1.67	
Общая	158	100	$N-1=17$		

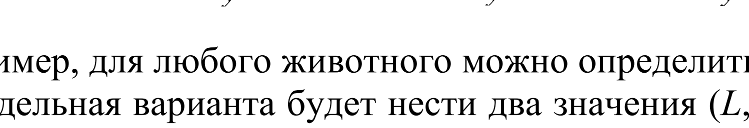
Таблица двухфакторного дисперсионного анализа имеет ту же структуру, что и таблица для однофакторного анализа, только факториальная дисперсия разложена на три компоненты (для факторов *A*, *B* и их взаимодействия). Для каждой из них требуется вычислить число степеней свободы с учетом числа градаций фактора *A* (*j*, количество столбцов) и числа градаций фактора *B* (*k*, количество рядов), значения дисперсий, а также критерий Фишера. Поскольку каждому из расчетных значений критерия соответствует свое число степеней свободы, табличные значения окажутся разными.

ОЦЕНКА ЗАВИСИМОСТИ МЕЖДУ ПРИЗНАКАМИ

Изложенные выше методы статистического анализа дают возможность изучать изменчивость биологических объектов по отдельным признакам – весу, размерам, плодовитости, физиологическим показателям и др. Однако в ряде случаев важно знать, какова зависимость между вариацией двух или нескольких признаков, изменяются ли две переменные самостоятельно, независимо друг от друга, или варьирование одного признака в какой-то степени связано с изменчивостью другого. В качестве второй переменной часто выступает какой-либо фактор среды.

Задачу исследования зависимостей можно рассматривать как развитие метода дисперсионного анализа, решающего задачу сравнения нескольких выборок, т. е. изучающего влияния фактора на признак. Техника дисперсионного анализа имеет две особенности. Фактор (или факториальный признак) задан дискретно, в виде градаций, или «доз». Когда исследуется фактор, заданный *качественно*, то разбиение на градации всего диапазона его действия оказывается очень эффективным способом создания подобия количественной переменной. Но при изучении количественно заданного фактора в грубой градуальной схеме дисперсионного анализа утрачивается часть информации, которая содержится в исходных выборках и которую можно было бы использовать. Кроме этого, дисперсионный анализ явным образом не учитывает тенденции изменения среднего уровня признака при изменении уровня фактора, не содержит показателя характера (знака) зависимости признака от фактора. Все эти «недостатки» дисперсионного анализа не характерны для методов изучения *сопряженной изменчивости* – корреляционного и регрессионного анализов.

Способ представления отдельных наблюдений здесь меняется: каждая варианта рассматривается как носитель двух численных характеристик объекта измерения, двух *зависимых* значений случайной величины. Если выше мы отождествляли отдельное значение с отдельной вариантой, то теперь мы рассматриваем варианту как некоторое тело, обладающее минимум двумя зарегистрированными качествами, различными у разных вариант:



Например, для любого животного можно определить массу (M) и длину (L) тела; отдельная варианта будет нести два значения (L, M). При этом множество вариант выборки можно отобразить графически как точки на плоскости осей двух признаков M и L. Вся выборка предстанет в виде множества точек на плоскости (двумерное рассеяние). Как видно на диаграмме (рис. 10), «облако» вариант вытянуто в направлении диагонали облака точек. Справа вверху находятся варианты с высокими значениями и размеров, и массы тела, в левом нижнем углу – с наименьшими значениями. В центре расположены варианты с промежуточными, средними значениями.

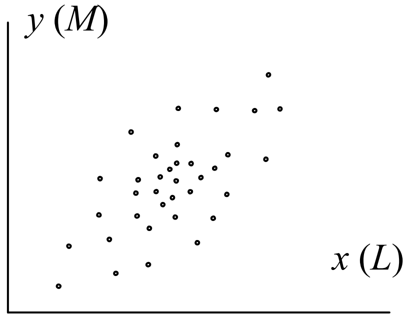


Рис. 10. *Область рассеяния вариант*

В первом приближении можно сказать, что *двумерное распределение – это ординация вариант на плоскости осей двух признаков*. Помимо рассеяния на плоскости, в определение двумерного распределения входит и частота встречаемости отдельных значений (*a*). Если признаки *x* и *y* теоретически подчиняются нормальному закону, тогда скопление вариант в трех осях (оси признаков *x*, *y* и частоты *a*) образует весьма странный «гребень», растянутое в пространстве *выпуклое нормальное распределение* (рис. 11). Однако в реальности такой идеальной картины получить никогда не удается, приходится ориентироваться только на плоскую фигуру рассеяния немногочисленных вариант. Если область, занятую вариантами, очертить по периферии плавной линией, мы получим вытянутую фигуру, эллипс, ограничивающий область рассеяния вариант, эллипс рассеяния. *Эллипс рассеяния – это область распространения вариант одной совокупности.*

Можно видеть, что в нашем примере признаки связаны друг с другом – есть общая тенденция: чем больше длина тела, тем больше вес; эта зависимость не очень жесткая, она размыта индивидуальными особенностями объектов (вариант).

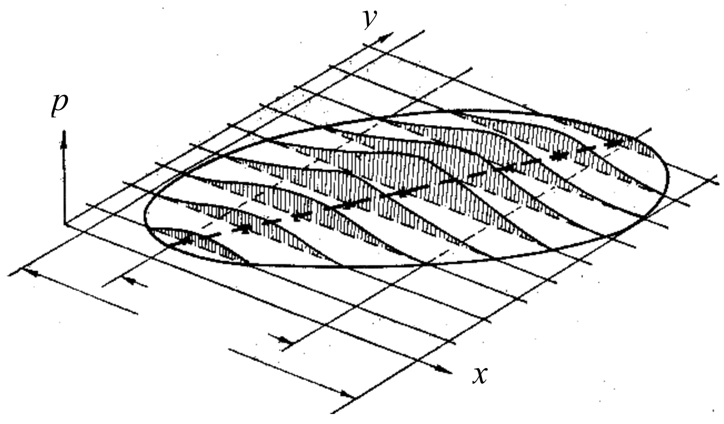


Рис. 11. *Двумерное распределение*

В двумерном распределении проявляются два эффекта: синхронное изменение двух признаков и размывание этой синхронности, т. е. действие факторов сопряжения признаков вдоль оси эллипса и действие случайных факторов – поперек нее.

Корреляционный анализ

Взаимная связь (взаимная зависимость) двух признаков при их изменчивости, т. е. сопряженность их вариации, называется корреляцией. Корреляция имеет место в тех случаях, когда признаки изменяются не автономно, а согласованно. Если с увеличением одного признака происходит соответствующее увеличение другого, говорят о положительной корреляции, и коэффициент корреляции имеет в этом случае положительный знак (+). Если же по мере увеличения первого признака второй уменьшается, то это отрицательная корреляция, коэффициент корреляции пишется со знаком минус (−). Полная положительная корреляция выражается единицей $r = 1$, полная отрицательная – $r = -1$. В природе такая ситуация встречается редко, и степень связи выражается той или иной долей единицы. При этом о тесной (сильной) корреляции обычно говорят в тех случаях, когда коэффициент корреляции не ниже ±0.6; значения ниже ±0.6 указывают на среднюю связь, а ниже ±0.3 – на слабую.

Коэффициент корреляции призван численно выражать долю сопряженной вариации двух признаков в общей их вариации:

$$r = \sqrt{\frac{\textit{ковариация}}{\textit{изменчивость}}} = \frac{Cxy}{\sqrt{Cx \cdot Cy}} = \frac{\sum (y - M_y)(x - M_x)}{\sqrt{\sum (y - M_y) \cdot \sum (x - M_x)}},$$

где Cxy – характеристика сопряженной изменчивости признаков,

Cx, Cy – характеристика общей изменчивости признаков.

При большом количестве данных коэффициент корреляции имеет смысл вычислять на компьютере (например, с помощью функции КОРРЕЛ в среде программы Excel), но для небольших выборок его можно быстро найти и при ручном счете. Рабочая формула для расчетов имеет вид:

$$r = \frac{Cxy}{\sqrt{Cx \cdot Cy}} = \frac{\sum xy - (\sum x \cdot \sum y)/n}{\sqrt{(\sum x^2 - (\sum x)^2/n) \cdot (\sum y^2 - (\sum y)^2/n)}}.$$

Способ вычисления коэффициента корреляции показан в таблице 13 на примере зависимости между живым весом коров (x) и их приплода (y, кг). По таблице рассчитываются квадраты вариант и их произведения, а также суммы вариант, квадратов и произведений. Вычисления ведутся по точным рабочим формулам.

Таблица 13

i	y	x	y^2	x^2	$x \cdot y$
1	25	352	625	123904	8800
2	26	376	676	141376	9776
3	31	402	961	161604	12462
4	32	453	1024	205208	14496
5	34	484	1156	234256	16456
6	38	528	1444	278784	20064
7	38	555	1444	308025	21090
Σ	224	3150	7330	1453158	103144

Проведем последовательные расчеты. Сначала определим вспомогательные величины:

$Cxy = \Sigma(x \cdot y) - (\Sigma x) \cdot (\Sigma y) / n = 103144 - 3150 \cdot 224 / 7 = 2344$,

$Cy = \Sigma y^2 - (\Sigma y)^2 / n = 7330 - 224^2 / 7 = 162$,

$Cx = \Sigma x^2 - (\Sigma x)^2 / n = 1453158 - 3150^2 / 7 = 35658$;

затем – коэффициент корреляции:

$$r = \frac{Cxy}{\sqrt{Cx \cdot Cy}} = \frac{2344}{\sqrt{35658 \cdot 162}} = 0.975.$$

Далее найдем его ошибку:

$$m_r = \sqrt{\frac{1 - r^2}{n - 2}} = \sqrt{\frac{1 - 0.975^2}{7 - 2}} = 0.099,$$

и, наконец, критерий t Стьюдента для проверки значимости коэффициентов:

$t_r = r / m_r = 0.975 / 0.099 = 9.84$.

Нулевая гипотеза предполагает отсутствие связи: «коэффициент корреляции значимо от нуля не отличается», $r = 0$. В нашем примере для уровня значимости $\alpha = 0.05$ и числа степеней свободы $df = n - 2 = 5$ находим табличное значение критерия Стьюдента $t_{(0.05,\,5)} = 2.57$. Полученная величина (9.84) значительно превышает табличную (2.57), что говорит о высокой статистической значимости коэффициента корреляции, о достоверности его отличия от нуля. Признаки *положительно* коррелируют, масса тела теленка действительно *возрастает* вслед за ростом массы тела коровы.

Выборный коэффициент корреляции в той или иной степени соответствует генеральному параметру. Определить диапазон, где лежит генеральное значение, можно с помощью доверительного интервала, хотя его *нельзя* построить непосредственно по формуле $r \pm t_{(\alpha,\,df)} \cdot m_r$. Дело в том, что область изменений коэффициента ограничена рамками ± 1, поэтому распределение выборочных коэффициентов корреляции в общем не соответствует нормальному (с диапазоном изменчивости $\pm\infty$). Поэтому перед расчетом коэффициент корреляции преобразуют в величину z, имеющую нормальное распределение, и уже для нее отыскивают границы доверительного интервала, после чего выполняют обратное преобразование.

Доверительный интервал для нашего случая ($r = 0.975$, $\alpha = 0.05$, $n = 7$, $df = n - 2 = 5$, $t_{(0.05,5)} = 2.57$) рассчитывается так. Преобразуем r:

$$z = 0.5 \cdot \ln\left(\frac{1 + r}{1 - r}\right) = 0.5 \cdot \ln\left(\frac{1 + 0.975}{1 - 0.975}\right) = 2.184$$

или берем его более точное значение из таблицы 13*П*, тогда $z = 2.0923$.

Определяем ошибку $m_z = \sqrt{\frac{1}{n - 3}} = \sqrt{\frac{1}{7 - 3}} = 0.5.$

Находим верхнюю границу: ${}_{\max}z = z + t_{(\alpha,df)} \cdot m_z = 2.09 + 2.57 \cdot 0.5 = 3.375$ и нижнюю границу: ${}_{\min}z = z - t_{(\alpha,df)} \cdot m_z = 2.09 - 2.57 \cdot 0.5 = 0.805$.

Обратное преобразование (по табл. 14*П*) дает: ${}_{\max}r \approx 1.00$, ${}_{\min}r \approx 0.67$. Истинное значение коэффициента корреляции находится в диапазоне от 0.67 до 1.00.

Ложная корреляция

Когда величина коэффициента корреляции определяется в первую очередь способом подбора вариант в выборку, а не реальной зависимостью между изучаемыми признаками, то говорят о «ложной корреляции».

Величина коэффициента корреляции зависит от вытянутости эллипса рассеяния: чем больше длина главной оси эллипса отличается от сечения, тем выше значение коэффициента. Случайные единичные, а тем более парные значения могут резко повысить показатель силы связи признаков. Особенно чувствителен коэффициент корреляции к нулям, которые могут попасть в исходную матрицу при переносе данных между электронными таблицами.

Явление ложной корреляции возникает и в том случае, когда исследуемые показатели имеют в сумме постоянное значение, например 100%. Рассмотрим соотношение численности грызунов и насекомоядных в разных биотопах (табл. 14). Представители и первого, и второго отрядов чаще встречаются в коренных хвойных лесах, нежели в антропогенных стациях, тем более в агроценозах. Синхронность их реакции на трансформацию ландшафтов выражается высоким коэффициентом корреляции их численности $r = 0.85$.

Таблица 14

Биотоп	Численность (экз./100 конусо-суток)			Доля, P (%)		
	бурозубок Nб	грызунов Nг	общая Nо	бурозубок Nб/ Nо	грызунов Nг / Nо	общая Nо / Nо
Кедровник	25	29	54	0.46	0.54	1
Смешанный	25	32	57	0.44	0.56	1
Экотон	23	21	44	0.52	0.48	1
Сосняк	22	16	38	0.58	0.42	1
Березняк	20	23	43	0.47	0.53	1
Луг	10	9	19	0.53	0.47	1
r	0.85			−1.00		

Если же оценить зависимость между долей грызунов (Pг = Nг/Nо) и долей бурозубок (Pб = Nб/Nо) в этих стациях (между индексами доминирования), она составит $r = -1.00$. Дело в том, что эти показатели рассчитываются относительно общей суммы, поэтому доля полевок составляет разницу между 1 и долей бурозубок: Pг = 1−Pб. По существу, мы имеем уравнение строго функциональной обратной регрессии ($y = 1 - 1 \cdot x$), которому соответствует, естественно, максимальный отрицательный коэффициент корреляции. Требование неизменности суммы двух показателей (1 или 100%), принятое для вычисления процентов, оказывается причиной постоянной обратной пропорции между этими показателями. Такая корреляция должна быть названа ложной, потому что характеризует не биологическую зависимость показателей, а способ их расчета. Когда общую сумму образуют три и более признаков, ложная корреляция будет отличаться от $r = -1$, но от этого не утратит своей природы математического артефакта.

При обработке массивов данных с большим числом производных признаков (индексы доминирования видов в сообществе, морфофизиологические индикаторы) нетрудно пропустить еще один вид ложной корреляции, которая наблюдается между двумя признаками, отнесенными к общей для них третьей переменной. По неосмотрительности коэффициенты связи между индексами могут быть восприняты как оценка зависимости между признаками. Такие корреляции, бессознательно наведенные третьим фактором, по сути являются ложными.

Безусловно, содержательную интерпретацию можно дать как корреляции признаков, так и корреляции индексов, но они будут кардинально отличаться. Например, среди нескольких видов куньих (от ласки до барсука) коэффициент корреляции между длиной тонкого и толстого отделов кишечника ($r = 0.96$) отражает простые морфологические пропорции: у крупного животного кишечник длиннее, чем у мелкого. Однако корреляция между индексами этих органов (размеров, отнесенных к длине тела особи) характеризует уже отличия диеты разных видов ($r = 0.78$): кишечник относительно меньше у облигатных хищников, нежели у полифагов. Однако в большом массиве производных значений такие отношения между индексами могут восприниматься как зависимости между признаками, что неизбежно приведет к ложным выводам.

Чтобы уйти от подобной двусмысленности, к обработке желательно привлекать только предварительно выверенные реальные исходные показатели, а не связанные методом расчета доли, проценты или индексы.

Множественная корреляция

Разобранные выше примеры корреляционных зависимостей касались главным образом взаимосвязи двух сопряженных процессов, явлений или варьирующих признаков. Между тем в практике биологических исследований нередко приходится сталкиваться с более сложными случаями, например, когда сопряжены не два, а три или более изменчивых фактора (признака). В такой ситуации возникает необходимость изучить множественные связи между большим числом взаимодействующих переменных, выступающих как в виде целой системы коррелированных признаков организма, так и в форме совместного влияния сложной совокупности факторов на определенное явление. Корреляционная зависимость нескольких переменных носит название множественной корреляции и оценивается коэффициентом, определяемым на основе корреляций между всеми парами признаков. Например, коэффициент множественной корреляции между тремя признаками *A*, *B* и *C* вычисляется по формуле:

$$r_{A.BC} = \sqrt{\frac{r_{AB}^2 + r_{AC}^2 - 2 \cdot r_{AB} \cdot r_{AC} \cdot r_{BC}}{1 - r_{AB}^2}}\ .$$

Полученная величина характеризует связь первого признака (*A*) с двумя другими (*B* и *C*). Покажем этот способ на примере совокупного действия двух факторов, *B* и *C* (температуры и влажности), на суточную активность

травяных лягушек (A). Определение парных корреляций дало следующие результаты (n = 110): $r_{AB} = +0.58$; $r_{AC} = +0.80$; $r_{BC} = -0.45$. Отсюда

$$r_{A.BC} = \sqrt{\frac{0.58^2 + 0.8^2 - 2 \cdot 0.58 \cdot 0.8 \cdot 0.45}{1 - 0.45^2}} = 0.86.$$

Сводный коэффициент корреляции оказался довольно высоким и, как показывает его сопоставление со стандартным значением по таблице 15*П*, вполне достоверным (при $\alpha < 0.001$).

С другой стороны, если обнаружена значительная корреляция между признаками A и C и между B и C, то не исключена возможность мнимой корреляционной зависимости между A и B, которая создается за счет одновременного влияния на них третьего признака C. Например, установленная по исследованиям в Карелии корреляция между численностью лесных полевок и урожаем семян сосны, скорее всего, объясняется не значением последних в питании грызунов (т. е. прямой причинной связью), а тем, что оба эти явления (численность полевок и урожай семян) контролируются одними и теми же экологическими факторами (прежде всего метеорологическими) и поэтому изменяются параллельно, хотя непосредственно между собой не связаны.

В этом и подобных случаях (например, когда настоящие зависимости между признаками животных маскируются влиянием возраста или когда связи между отдельными промерами организма создаются за счет влияния живого веса и т. д.) возникает задача изучить корреляцию между двумя признаками (A и B), исключив влияние на эту связь третьего признака (C), как бы элиминировав его.

Частная корреляция

Этой цели служит коэффициент частной корреляции, оценивающий связь между первым и вторым признаками при постоянных значениях третьего и вычисляемый по формуле:

$$r_{A(BC)} = \frac{r_{AB} - r_{AC} \cdot r_{BC}}{\sqrt{(1 - r_{AC}^2) \cdot (1 - r_{BC}^2)}},$$

где A и B – факторы, связь которых требуется изучить;

C – фактор, влияние которого необходимо исключить из корреляционной зависимости между A и B (реперный признак);

r_{AB}, r_{AC}, r_{BC} – соответствующие парные коэффициенты корреляции, вычисляемые обычным способом;

$r_{A(BC)}$ – искомый коэффициент частной корреляции, показывающий связь между двумя признаками при исключении влияния третьего.

Этот же метод можно применить и для элиминации двух факторов при четырех переменных и т. д. Формула для расчетов примет в этом случае следующий вид:

$$r_{AB(BD)} = \frac{r_{AB(C)} - r_{AC(B)} \cdot r_{BC(D)}}{\sqrt{(1 - r^2_{AC(D)}) \cdot (1 - r^2_{BC(D)})}}.$$

Рассмотрим нахождение коэффициента частной корреляции на упрощенном примере (взятом из книги П. Ф. Рокицкого). Получены данные о корреляции между давлением крови (*A*), содержанием в ней холестерина (*B*) и возрастом (*C*) у 142 женщин. Соответствующие коэффициенты корреляции таковы: $r_{AB} = +0.25$; $r_{AC} = +0.33$; $r_{BC} = 0.51$.

Известно, что повышенное артериальное давление может быть связано с высоким содержанием холестерина в стенках кровеносных сосудов, однако и давление крови, и концентрации холестерина увеличиваются с возрастом. Поэтому возникает вопрос, создается ли корреляция между давлением крови и содержанием в ней холестерина за счет их общей связи с возрастом или же она реально существует для каждого возраста (и независимо от него). Элиминируя эффект возраста по приведенной выше формуле, получим:

$$r_{A(BC)} = \frac{0.25 - 0.33 \cdot 0.51}{\sqrt{(1 - 0.33^2) \cdot (1 - 0.5^2)}} = 0.12.$$

По таблице 15*П* можно установить, что при $n = 150$ для достоверности коэффициента корреляции даже при уровне значимости $\alpha = 0.05$ его величина должна быть не меньше 0.159. В данном же случае полученное значение меньше табличного и, следовательно, коэффициент корреляции от нуля достоверно не отличается. Таким образом, внутри отдельных возрастных групп корреляционной связи между давлением крови и содержанием холестерина, по крайней мере на изученном материале, не обнаруживается. Пока нет оснований отбрасывать нулевую гипотезу.

Второй пример демонстрирует использование коэффициента частной корреляции для более глубокого проникновения в структуру нескольких факторов наведения. Рассмотрим выборку объектов разного статуса (11 видов мелких млекопитающих), взяв в качестве признаков их численность в семи биотопах прибайкальской равнины. Реперным признаком послужила суммарная численность вида во всех биотопах. Здесь коэффициент корреляции отражает сходство между биотопами по соотношениям численности 11 видов. Например, оказалось, что между березняком и экотоном (граница между березняком и коренными лесами) и общая корреляция ($r = 0.92$), и частная ($r = 0.64$) высока и положительна. Можно утверждать, что население животных этих биотопов почти идентично.

В свою очередь, корреляция между кедровником и лугом не проявилась ($r = -0.08$), но коэффициент частной корреляции был велик и отрицателен ($r = -0.43$). Этим оттеняется тот факт, что виды, отсутствующие на лугу, многочисленны в кедровнике (красная полевка, мышь), а обычные в агроценозе – крайне редки в тайге (серые полевки).

Частная корреляция показала, что население этих биотопов во многом диаметрально противоположно. Она выявила два вида факторов наведения. Один из них хорошо известен – это сезонное расселение видов в другие био-

топы. В течение периода размножения видовой состав тайги и луга меняется несогласованно (одни виды идут из тайги в агроценозы, другие – в противоположном направлении) и численность всех видов относительно выравнивается, $r = -0.08$. Частная корреляция устраняет эффект прироста численности за счет иммигрантов и выдвигает на первый план контраст «базовой» численности, которую формируют характерные обитатели биотопов: в тайге это лесные полевки, на лугу – серые. Так *проявляется* второй фактор: отличие качества среды в разных биотопах. Он обеспечивает формирование принципиально несходных зооценозов, что и показывает высокой частной корреляцией $r = -0.43$.

Ранговая корреляция

Помимо рассмотренных выше параметрических показателей связи в биометрии применяются и непараметрические. Обычно их используют при сильных отклонениях изучаемого распределения от нормального (или сомнениях на этот счет), а также в тех случаях, когда требуется оценить зависимость между качественными или полуколичественными признаками, точное количественное измерение которых затруднено (оценки в баллах или других условных единицах). Если варианты выборки могут быть упорядочены по степени выраженности их свойств, для измерения степени сопряженности между ними можно воспользоваться *непараметрическим показателем связи* – ранговым коэффициентом корреляции Спирмена:

$$r_S = 1 - \frac{6 \cdot \sum d^2}{n \cdot (n^2 - 1)},$$

где d – разность между рангами сопряженных значений признаков x и y;
n – объем выборки.

Этой формулой следует пользоваться в тех случаях, когда выборки не содержат повторяющихся вариант, когда все ранги выражены разными целыми числами. Если же исходные ряды содержат одинаковые значения, расчет корреляции приходится вести по другой формуле, включающей поправку на повторы (при этом одинаковым вариантам присваивается средний ранг):

$$r_S = \frac{\frac{(n^3 - n)}{6} - (T_x + T_y) - \sum d^2}{\sqrt{\left(\frac{(n^3 - n)}{6} - 2 \cdot T_x\right)\left(\frac{(n^3 - n)}{6} - 2 \cdot T_y\right)}},$$

где T_x, T_y – поправки на серии повторов для каждой выборки:

$$T_x = \frac{\sum^{k} (t_x^{\,3} - t_x)}{12},$$

где t – число членов в каждой группе одинаковых вариант.

Поправки Tx, Ty учитывают k групп повторяющихся вариант.

Рассмотрим технику вычислений на примере изучения связи между оцененными в баллах численностью лисицы (x) и обилием мышевидных грызунов (y) (по годам наблюдений):

	1957	1958	1959	1960	1961	1962	1963	1964	1965	1966
x	2.6	2.1	2.3	2.3	1.6	2.2	3.0	2.1	1.5	2.2
y	3.0	2.4	3.6	2.9	3.7	3.3	4.0	2.1	1.0	3.5

Чтобы проверить наличие и определить силу этой связи, нужно упорядочить значения сопряженных признаков по степени их выраженности, затем присвоить им ранги, обозначив значения порядковыми числами натурального ряда, и рассчитать коэффициент корреляции. Техника вычислений показана в таблице 15.

Таблица 15

Численность лисицы в баллах, x	Обилие грызунов в баллах, y	Ранги вариант		Разность между рангами, d	d^2
		R_x	R_y		
1.5	1.0	1	1	0	0
1.6	3.7	2	6	−4.0	16.00
2.1	2.4	3.5	3	+0.5	0.25
2.1	2.1	3.5	2	+1.5	2.25
2.2	3.3	5.5	7	−1.5	2.25
2.2	3.6	5.5	8.5	−3.0	9.00
2.3	3.6	7.5	8.5	−1.0	1.00
2.3	2.9	7.5	4	+3.5	12.25
2.6	3.0	9	5	+4.0	16.00
3.0	4.0	10	10	0	0
					Σ = 59

В ряду значений признака x есть три пары одинаковых вариант, поэтому поправка будет равна: $T_x = \dfrac{(2^3-2)+(2^3-2)+(2^3-2)}{12} = 1.5$.

В ряду признака y всего одна пара одинаковых значений; поправка составит: $T_y = \dfrac{(2^3-2)}{12} = 0.5$.

Находим величину $\dfrac{(n^3-n)}{6} = \dfrac{(10^3-10)}{6} = 165$.

Коэффициент ранговой корреляции составит:

$$r_S = \frac{165-(1.5+0.5)-59}{\sqrt{(165-2\cdot1.5)(165-2\cdot0.5)}} = 0.638.$$

Если воспользоваться формулой без поправок, результат будет несколько иным:

$$r_S = 1 - \frac{6 \cdot \sum d^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot 59}{10 \cdot (10^2 - 1)} = 0.642.$$

Статистическая ошибка и критерий достоверности отличия коэффициента корреляции от нуля вычисляются по формулам:

$$m_r = \sqrt{\frac{1 - r_S^2}{n - 2}} = \sqrt{\frac{1 - 0.638^2}{10 - 2}} = 0.272,$$

$$t_r = r_S \,/\, m_r = 0.638 \,/\, 0.272 = 2.34.$$

Величина критерия (2.34) несколько выше критического значения (2.31) для уровня значимости $\alpha = 0.05$ и числа степеней свободы $df = n - 2 = 8$ (табл. 6*П)*. Казалось бы, это дает основание отвергнуть нулевую гипотезу ($r_S = 0$) и с вероятностью $P = 95\%$ констатировать достоверность установленной связи. Однако при небольших выборках статистические свойства коэффициента Спирмена не очень «хороши» и для оценки значимости корреляции лучше воспользоваться специально подготовленной таблицей 16*П*, аналогичной рассмотренной выше таблице 15*П*.

Чтобы полученный коэффициент можно было считать достоверно отличным от нуля, он должен превышать табличное значение при данном n. В нашем случае ($n = 10$, $\alpha = 0.05$) коэффициент $r = 0.638$ ниже табличного $r = 0.64$, следовательно, значимо от нуля не отличается. Зависимость численности лисицы и грызунов по приведенным данным достоверно не прослеживается.

Коэффициент контингенции

Степень сопряженности (сочетаемость) двух возможных состояний двух качественных признаков также можно измерить с помощью особого коэффициента корреляции – коэффициента контингенции Шарлье.

У каждой особи отмечают два признака, имеющих альтернативные распределения, и вся выборка разбивается на четыре части:

a – число особей, имеющих оба признака (+ +),

b – число особей, имеющих первый признак, но не имеющих второго (+ −),

c – число особей, не имеющих первого признака, но имеющих второй (− +),

d – число особей, не имеющих обоих признаков (− −).

На схеме это выглядит как четырехклеточная корреляционная решетка:

Признак 2 / Признак 1	Присутствует (+)	Отсутствует (−)	Σ
Присутствует (+)	*a*	*c*	*a* + *c*
Отсутствует (−)	*b*	*d*	*b*+ *d*
Σ	*a* + *b*	*c* + *d*	*n* = *a* + *b* + *c* + *d*

Степень взаимосвязи определяется по формуле:

$$r = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b)\cdot(c+d)\cdot(a+c)\cdot(b+d)}}.$$

При вычислении коэффициента корреляции между двумя альтернативными признаками выясняется вопрос о том, чаще ли оба признака одновременно присутствуют или отсутствуют у варианты, чем это могло бы быть по случайным причинам. Достоверность отличия от нуля оценивается по критерию Стьюдента: $t_r = r / m_r$, где $m_r = \frac{1 - r^2}{\sqrt{n-1}}$.

При проверке влияния перекрытий на оплодотворяемость самок песцов получены первичные материалы о численности родивших (+) и неродивших (−) самок из числа хотя бы дважды перекрытых (+) и неперекрытых (−).

Признак 2 / Признак 1	Родившие (+)	Неродившие (−)	Σ
Перекрытые (+)	370	90	460
Неперекрытые (−)	100	120	220
Σ	470	210	$n = 680$

Коэффициент ассоциации равен: $r = \frac{370 \cdot 120 - 100 \cdot 90}{\sqrt{470 \cdot 210 \cdot 460 \cdot 220}} = 0.35.$

Ошибка коэффициента корреляции составит: $m_r = \frac{1 - 0.35^2}{\sqrt{680 - 1}} = 0.0327,$

а критерий Стьюдента $t_r = 0.35 / 0.0327 = 10.7$.

Полученное значение (10.7) настолько велико, что превышает табличное даже для доверительной вероятности выше $P = 0.999$ (уровень значимости $\alpha < 0.001$). Влияние повторных покрытий на оплодотворяемость самок песцов несомненно.

При исследовании связи между белой мастью и красными глазами у кроликов получены следующие данные.

Глаза / Шерсть	Красные	Некрасные	Σ
Белая	29	11	40
Окрашенная	1	59	60
Σ	30	70	100

Подстановка всех значений сумм из таблицы в формулы дает: $r = 0.76$, $m = 0.04$, $t = 19$. Достоверность связи не вызывает сомнений.

Регрессионный анализ

Коэффициент корреляции указывает лишь на степень (тесноту) связи в изменчивости двух переменных величин, но не позволяет судить о том, как меняется одна величина по мере изменения другой. Ответ на этот вопрос дает вычисление коэффициента регрессии, показывающего, на какую величину в среднем изменяется один признак при изменении другого на единицу измерения. Регрессионный анализ, в отличие от корреляционного, изучает эффект *влияния одного признака на другой*, зависимость признака от фактора, характер влияния фактора на признак. Его основные результаты таковы:

1. Таблица дисперсионного анализа, в которой показаны сила и достоверность влияния на признак изучаемого фактора или другого признака.

2. Уравнение регрессии, выражающее пропорциональность сопряженного изменения признаков, тенденции их взаимосвязанной изменчивости или динамики.

3. Оценки значимости коэффициентов уравнения регрессии.

Регрессионный анализ методически ориентирован односторонне – на изучение зависимости одного признака от другого (зависимость y от x или, напротив, зависимость x от y), хотя может применяться к случаям, когда фактически имеется взаимозависимость двух переменных.

Основную тенденцию взаимосвязанного изменения двух признаков можно отобразить с помощью простого графического приема. Разобьем ось x на несколько интервалов. Найдем для каждого из них частные средние значения признака y (M_y). Теперь проведем через эти средние точки ломаную линию. Это будет линия регрессии Y по x. *Регрессия – изменение среднего уровня одного признака при изменении другого* (рис. 12).

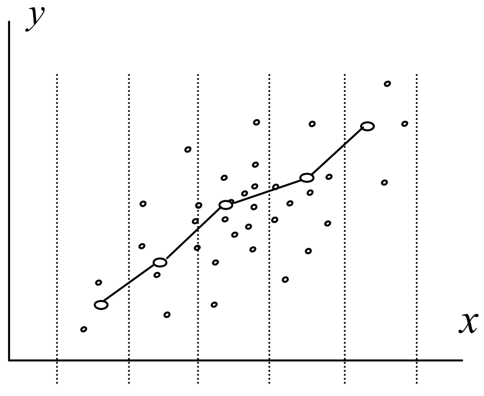


Рис. 12. *Эмпирическая линия регрессии*

Линейная регрессия

К сожалению, ход ломаной линии нельзя передать простым уравнением, к тому же на нем сказываются способ интервального разбиения оси абсцисс, а также уровень репрезентативности в разных областях распределения. В этом смысле предпочтительнее единственная прямая линия регрессии, подчеркивающая основные тенденции зависимости признаков, которая может быть выражена простым уравнением линии: $y = ax + b$.

Судить о том, как меняется одна величина по мере изменения другой, позволяет коэффициент регрессии (a), показывающий, на какую величину в среднем изменяется один признак (y) при изменении другого (x) на единицу измерения (точнее, на какую величину один признак отклоняется от своей средней при некотором отклонении другого признака от своей средней):

$y - My = a \cdot (x - Mx)$.

Простые преобразования:

$y = a \cdot x + My - a \cdot Mx$, $b = My - a \cdot Mx$

и приводят к уравнению линии: $y = ax + b$.

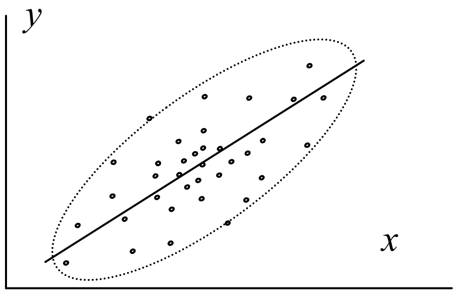


Рис. 13. *Линейная регрессия*

Рассчитать коэффициенты уравнения регрессии позволяет *метод наименьших квадратов*, основная идея которого состоит в том, чтобы линия регрессии прошла на наименьшем удалении от каждой точки, т. е. чтобы сумма квадратов расстояний от всех точек до прямой линии была наименьшей. В математической статистике показано, что для случая двумерного нормального распределения лучшей (эффективной, несмещенной и пр.) линией, описывающей зависимость одного признака от другого, может быть только линия частных средних арифметических.

Вычисления коэффициентов линейной регрессии $y = ax + b$ ведутся по следующему алгоритму. Сначала найдем вспомогательные величины:

$Cx = \Sigma x^2 - (\Sigma x)^2 / n$,

$Cy = \Sigma y^2 - (\Sigma y)^2 / n$,

$Cxy = \Sigma(x \cdot y) - (\Sigma x) \cdot (\Sigma y) / n$,

$M_y = \Sigma y / n$, $M_x = \Sigma x / n$.

Затем рассчитаем коэффициенты: $a = Cxy / Cx$, $b = M_y - a \cdot M_x$.

Оценить значимость коэффициента регрессии позволяет критерий t Стьюдента, проверяющий нулевую гипотезу Но: $a = 0$, коэффициент регрессии значимо от нуля не отличается. С этой целью рассчитывается ошибка коэффициента регрессии m_a:

$m_a = \frac{S_y}{S_x} \cdot m_r$, где m_r – ошибка коэффициента корреляции (см. с. 62),

и вычисляется значение критерия:

$t = (a - 0) / m_a = a / m_a \sim t_{(0.05,\, n-2)}$.

Смысл этого критерия состоит в следующем. Коэффициент регрессии a характеризует сопряженность пропорционального изменения двух признаков, т. е. отвечает за то, что линия регрессии имеет некоторый угол относительно оси абсцисс. Значение $a = 0$ означает, что линия регрессии идет параллельно оси OX, что при изменении признака x признак y не меняется, т. е.

что y не зависит от x. Значения коэффициента, отличные от нуля, говорят о том, что взаимосвязь признаков имеет место, при $a > 0$ зависимость положительная, при $a < 0$ – отрицательная.

Вернемся к примеру с описанием зависимости между живым весом коров и их приплода (стр. 61). Расчеты для построения уравнения регрессии показаны в таблице 16. Сначала вычисляются квадраты вариант и их произведения, а также суммы вариант, квадратов и произведений. Вычисления ведутся по точным рабочим формулам. Проще всего это делать в среде Excel, с помощью команды Сервис \ Анализ данных \ Регрессия.

Таблица 16

i	y	x	y^2	x^2	$x \cdot y$	Y	$(y-Y_i)^2$	$t \cdot m_Y$	min$Y$	max$Y$
1	25	352	625	123904	8800	25.6	0.31	2.0	23.6	27.5
2	26	376	676	141376	9776	27.1	1.29	1.7	25.5	28.8
3	31	402	961	161604	12462	28.8	4.65	1.4	27.4	30.2
4	32	453	1024	205208	14496	32.2	0.04	1.2	31.0	33.4
5	34	484	1156	234256	16456	34.2	0.06	1.3	32.9	35.5
6	38	528	1444	278784	20064	37.1	0.76	1.7	35.4	38.9
7	38	555	1444	308025	21090	38.9	0.81	2.1	36.8	41.0
Σ	224	3150	7330	1453158	103144		7.92			

Проведем последовательные расчеты вручную. Сначала определим вспомогательные величины:

$n = 7$,

$Cxy = \Sigma(x \cdot y) - (\Sigma x) \cdot (\Sigma y)/n = 103144 - 3150 \cdot 224 / 7 = 2344$,

$Cy = \Sigma y^2 - (\Sigma y)^2 / n = 7330 - 224^2 / 7 = 162$,

$Cx = \Sigma x^2 - (\Sigma x)^2 / n = 1453158 - 3150^2 / 7 = 35658$,

затем – параметры:

$M_y = \Sigma y / n = 224 / 7 = 32$,

$M_x = \Sigma x / n = 3150 / 7 = 450$,

$$S_y = \sqrt{\frac{Cy}{n-1}} = \sqrt{\frac{162}{6}} = 5.2,$$

$$S_x = \sqrt{\frac{Cx}{n-1}} = \sqrt{\frac{35658}{6}} = 77.1,$$

$$r = \frac{Cxy}{\sqrt{Cx \cdot Cy}} = \frac{2344}{\sqrt{35658 \cdot 162}} = 0.975,$$

$$a = \frac{Cxy}{Cx} = \frac{2344}{35658} = 0.0657,$$

$b = M_y - a \cdot M_x = 32 - 0.0657 \cdot 450 = 2.419$.

Получено уравнение линейной регрессии $Y = 0.0657x + 2.419$, которое позволяет рассчитать теоретические значения Y (табл. 16, графа 7).

Далее найдем ошибку коэффициента регрессии:

$$m_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0.975^2}{7-2}} = 0.099,$$

$$m_a = \frac{S_y}{S_x} \cdot m_r = \frac{5.2}{77.1} \cdot 0.099 = 0.00667,$$

и, наконец, критерий t Стьюдента для *проверки значимости коэффициента регрессии*: $t_a = a / m_a = 0.0657 / 0.00667 = 9.84$.

Для уровня значимости $\alpha = 0.05$ и числа степеней свободы $df = n - 2 = 5$ находим табличное значение критерия Стьюдента $t_{(0.05,5)} = 2.57$. Полученная величина (9.84) превышает табличную (2.57), что говорит о статистической значимости коэффициента регрессии (a), о достоверности его отличия от нуля. Масса тела теленка действительно возрастает вслед за ростом массы тела коровы.

Рассчитаем *доверительную зону* (интервал), *в которой с той или иной вероятностью заключены теоретические средние значения* веса новорожденных. Критерий Стьюдента (нормированное отклонение) для уровня значимости $\alpha = 0.05$, и числа степеней свободы $df = n - 1 = 6$ составит 2.45. Далее находим границы. Так, для значения $x = 352$ кг прогноз по уравнению регрессии равен $Y = 25.56$, а возможное отклонение средней составит:

$$t \cdot m_Y = t \cdot m_y \cdot \sqrt{\frac{1}{n} + \frac{(x_i - M_x)^2}{Cx}} = 2.45 \cdot 1.2582 \cdot \sqrt{\frac{1}{7} + \frac{(352 - 450)^2}{35658}} =$$

$$= 2.45 \cdot 0.81 = 1.98.$$

Отсюда находим границу доверительного интервала (табл. 16):
верхнюю: $\max Y = Y_i + t \cdot m_Y = 25.56 + 1.98 = 27.54$
и нижнюю: $\min Y = Y_i - t \cdot m_Y = 25.56 - 1.98 = 23.58$.

Средняя масса новорожденного теленка для коров весом 352 кг с вероятностью $P = 0.95$ должна находиться в диапазоне от 23.6 до 27.5 кг (рис. 14).

Регрессионный анализ позволяет проверить *значимость* и второго коэффициента уравнения регрессии, *свободного члена b*. Математический смысл свободного члена уравнения линии состоит в том, что этому значению равна функция (y) при условии, что аргумент равен нулю ($x = 0$):

$$y = ax + b = a \cdot 0 + b = b.$$

В рамках регрессионного анализа рассматривается именно эта гипотеза Но: $b = 0$, т. е. что линия регрессии проходит через начало осей координат, точку пересечения осей координат, через нуль. Если гипотеза опровергается, значит, линия регрессии не пересекает ось ординат. Если гипотеза не опровергается, мы можем считать, что между признаками существует простая пропорция ($Y = ax$) и расчет коэффициента регрессии a упрощается: $a = \Sigma(x \cdot y) / \Sigma x^2$. Нулевая гипотеза Но: $b = 0$ проверяется по критерию Стьюдента: $t = (b - 0) / m_b = b / m_b \sim t_{(0.05,\, n-2)}$, где m_b – ошибка коэффициента b.

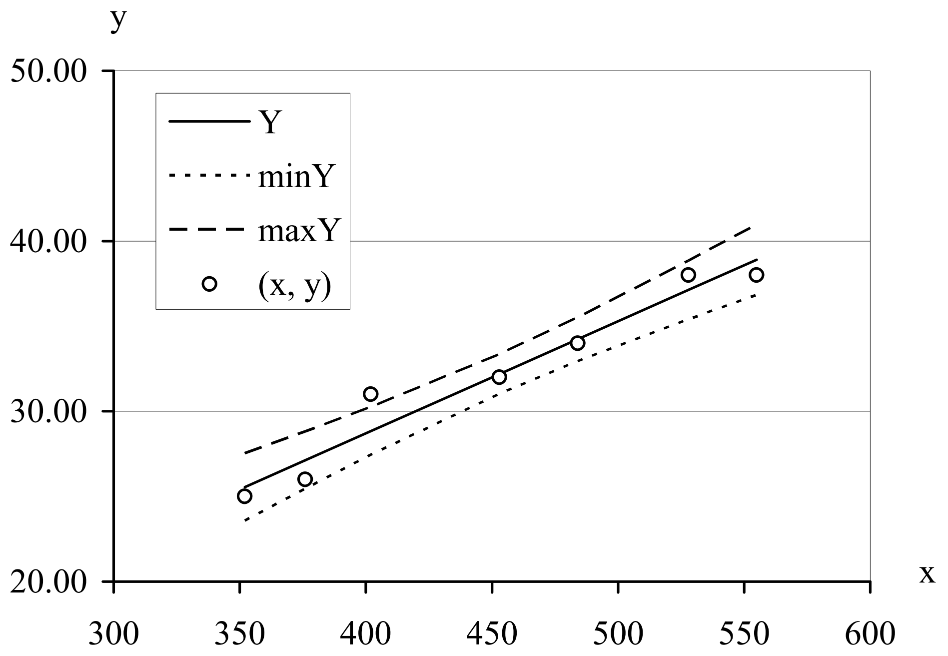


Рис. 14. *Линия регрессии Y = 0.0657·x+ 2.1347 и ее доверительный интервал*

Ошибка второго коэффициента регрессии рассчитывается в два этапа. Сначала находим общую ошибку регрессионной средней (или остаточное стандартное отклонение), которая может вычисляться по-разному.

Точная формула *для небольших выборок* дает величину:

$$m_y = S_y \cdot \sqrt{\frac{(n-1)\cdot(1-r^2)}{n-2}} = 5.2\cdot\sqrt{\frac{(7-1)\cdot(1-0.975^2)}{7-2}} = 1.2582.$$

Общая точная формула показывает практически такой же результат:

$$m_y = \sqrt{\frac{\sum^{n}(y_i - Y_i)^2}{n-2}} = \sqrt{\frac{C_{остат.}}{n-2}} = \sqrt{S^2_{остат.}} = \sqrt{\frac{7.92}{5}} = \sqrt{1.5832} = 1.2582$$

(величина $C_{остат.} = \sum^{n}(y_i - Y_i)^2$ – это сумма квадратов разности между расчетными и реальными значениями признака, она найдена в табл. 16, внизу 7 графы, $C_{остат.}$= 7.92).

Теперь вычисляем ошибку коэффициента *b*:

$$m_b = m_y \cdot \sqrt{\frac{1}{n} + \left(\frac{M_x}{Cx}\right)^2} = 1.2582\cdot\sqrt{\frac{1}{7} + \left(\frac{450}{35658}\right)^2} = 3.0359$$

и критерий *t* Стьюдента: $t_b = b\,/\,m_b$ = 2.419 / 3.0359 = 0.797.

Для уровня значимости α = 0.05 и числа степеней свободы $df = n - 2 = 5$ табличное значение составляет $t_{(0.05,\,5)}$ = 2.57. Анализ показал, что критерий Стьюдента для свободного члена уравнения (0.797) оказался ниже табличного значения (2.57), т. е. коэффициент *b* значимо от нуля не отличается (при данном объеме собранных материалов). Это позволяет пересчитать коэффи-

циент регрессии: $a = \Sigma(x \cdot y) / \Sigma x^2 = 0.071$. Теперь можно пользоваться уравнением регрессии вида: $Y = 0.071 \cdot x$.

Оценить *достоверности взаимодействия признаков* можно и с помощью дисперсионного анализа (табл. 17). В этом случае общая дисперсия зависимого признака y ($C_{общ.}$) разлагается на две составляющие – регрессионную дисперсию (изменчивость признака y, связанная с влиянием признака x ($C_{регр.}$), и случайную, или остаточную, дисперсию (изменчивость признака y, связанная с влиянием неучтенных случайных факторов ($C_{остат.}$) (рис. 14, табл. 17, 18).

Общую сумму квадратов ($C_{общ.} = C_y = \Sigma(y_i - M_y)^2 = \Sigma y_i^2 - (\Sigma y_i)^2 / n$) находят непосредственно как сумму квадратов отличий между значением y_i для каждой варианты и общей средней признака y. Остаточную сумму квадратов ($C_{остат.} = \Sigma(y_i - Y_i)^2$) находят также непосредственно как сумму квадратов отличий между значением y_i для каждой варианты и значением, предварительно рассчитанным по уравнению регрессии $Y_i = ax_i + b$ (для соответствующих значений x_i). Модельную сумму квадратов ($C_{мод.} = \Sigma(Y_i - M_y)^2$) рассчитывают как разность между общей и остаточной ($C_{мод.} = C_{общ.} - C_{остат.}$).

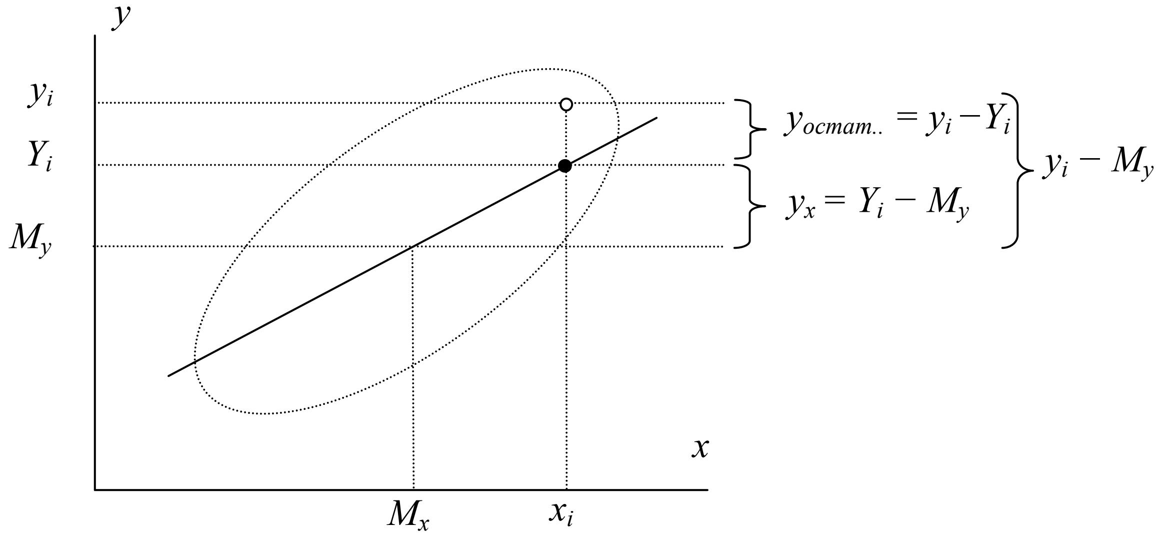


Рис. 15. *Модель варианты в регрессионном анализе*

Таблица 17

Составляющие дисперсии	Суммы квадратов, *C*	Формулы расчета сумм квадратов	*df*	S^2	*F*
Регрессия	$C_{регр.} = \Sigma(Y_i - M_y)^2$	$C_{общ.} - C_{остат.}$	1	$S^2_{регр.} = \dfrac{C_{регр.}}{df_{регр.}}$	$\dfrac{S^2_{регр.}}{S^2_{остат.}}$
Отклонения вариант от линии регрессии	$C_{остат.} = \Sigma(y_i - Y_i)^2$		$n - 2$	$S^2_{остат.} = \dfrac{C_{остат.}}{df_{остат.}}$	$F_{(0.05,\ 1,\ n-2)}$
Общая (всего)	$C_{общ.} = \Sigma(y_i - M_y)^2$	$(\Sigma y_i^2 - \Sigma y_i)^2 / n = C_y$			

Таблица 18

Составляющие дисперсии	C		df	S^2	F
Регрессия	$C_{регр.} = \Sigma (Y_i - Y)^2$	154.08	1	$S^2_{регр.} = 154.08$	$F = \frac{154.08}{1.58} = 97.3$
Отклонения вариант от линии регрессии	$C_{остат.} = \Sigma (y_i - Y_{xi})^2$	7.92	5	$S^2_{остат.} = 1.58$	$F_{(0.05,\,1,\,5)} = 6.6$
Общая (всего)	$C_{общ.} = \Sigma (y_i - Y)^2$	162			

Показателем «силы влияния признака на признак» служит *коэффициент детерминации, отношение регрессионной суммы квадратов к общей сумме квадратов* (принимает значения от 0 до 1): $R^2 = \frac{C_{мод.}}{C_{общ.}} = \frac{154.08}{162} = 0.95.$

Между коэффициентом детерминации и коэффициентом корреляции существует простое соответствие: $r = \sqrt{R} = \sqrt{0.95} = 0.975$.

Построив таблицу дисперсионного анализа с помощью критерия Фишера, можно проверить нулевую гипотезу Но: предсказания регрессионной модели в целом неадекватно описывают исходные данные, зависимости между признаками нет. Конструкция критерия исследует вопрос, превышает ли варьирование, учтенное моделью, случайное (остаточное) варьирование? Критерий Фишера вычисляется как отношение модельной и остаточной дисперсии:

$F = S^2_{мод.} / S^2_{остат.} = 154.08 / 1.58 = 97.3.$

Табличное значение $F_{(0.05,\,1,\,5)} = 6.6$. Поскольку полученное значение критерия оказалось выше табличного, дисперсия реального признака y приближается по величине к дисперсии расчетных значений признака Y, т. е. существенно превышает (случайные) отличия между ними. Регрессионная модель в целом адекватно описывает исходные данные.

Криволинейная регрессия

В большинстве случаев связь биологических признаков не бывает линейной, они изменяются либо с разной скоростью, либо в разных масштабах. Соответственно на графике форма такой связи отображается не прямой, а кривой линией. Примерами могут служить геометрическая прогрессия роста численности популяции в оптимальных условиях, различие скоростей роста разных частей тела, определяющее аллометрический характер зависимости признаков (лицевой отдел черепа растет более интенсивно, чем мозговой). В подобных случаях эффективнее использовать не уравнения прямой линии ($y = ax + b$), а разнообразные уравнения кривых линий, например, степенной, гиперболической, экспоненциальной, параболической, логистической и др.

Поскольку метод наименьших квадратов исходно ориентирован на линию (поиск уравнения линии, наименее удаленной ото всех эмпирических точек), прямой расчет уравнений кривых в рамках регрессионного анализа невозможен. Натурные данные необходимо предварительно «выпрямить», т. е. сделать возможным вычисление *линейного уравнения регрессии* с тем, чтобы потом из него получить уравнение криволинейной связи. Общий порядок регрессионного анализа для криволинейной зависимости следующий:

- преобразование исходных данных, «выпрямляющее» зависимость,
- расчет коэффициентов линейной регрессии преобразованных данных,
- проведение дисперсионного анализа, оценка значимости коэффициентов регрессии,
- обратное преобразование коэффициентов линейной регрессии для конструирования уравнения криволинейной регрессии.

Рассмотрим процесс поиска уравнения криволинейной регрессии на примере изучения зависимости веса печени прыткой ящерицы от длины ее тела (рис. 16).

Рассчитанное по исходным данным уравнение линейной регрессии имеет вид: $y = 107.9x - 404.2$. И хотя коэффициент регрессии достоверен ($t = 7.6$, $\alpha < 0.05$) и коэффициент детерминации высок $R^2 = 0.866$, это уравнение весьма приблизительно описывает зависимость признаков – для наименьших наблюдаемых значений длины тела оно дает абсурдное (отрицательное) значение массы печени ($107.9 \cdot 3.4 - 404.2 = -37.3$ мг). Линейная модель не годится даже для интерполяции изучаемых данных. Гораздо успешнее справляется с подобной задачей степенная (аллометрическая) функция $y = bx^a$.

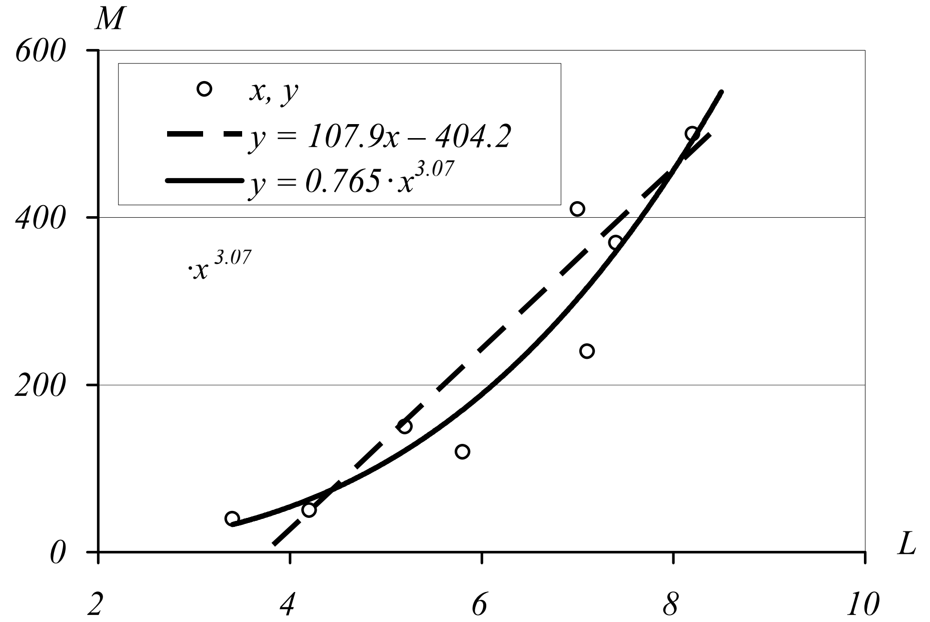


Рис. 16. *Зависимость веса печени (M, мг) от длины тела (L, мм) у ящериц*

Для вычисления коэффициентов этого уравнения воспользуемся преобразованием: $Y = \lg y$, $X = \lg x$, $B = \lg b$. После логарифмирования степенное уравнение приняло линейный вид: $\lg y = \lg b + a \cdot \lg x$ или $Y = B + aX$. Теперь остается отыскать коэффициенты уравнения B и a, используя алгоритм метода наименьших квадратов (табл. 19).

Таблица 19

№	x	y	$X = \lg x$	$Y = \lg y$	X^2	Y^2	X·Y	Y'	$(Y'-Y)^2$	y'
1	3.4	40	0.531	1.60	0.282	2.567	0.85	1.517	0.00718	33
2	4.2	50	0.623	1.69	0.388	2.886	1.06	1.799	0.01009	63
3	5.2	150	0.716	2.18	0.513	4.735	1.56	2.085	0.00838	121
4	5.8	120	0.763	2.08	0.583	4.323	1.58	2.23	0.02284	170
5	7.1	240	0.851	2.38	0.725	5.665	2.03	2.5	0.01442	316
6	7.0	410	0.845	2.61	0.714	6.827	2.21	2.481	0.01728	303
7	7.4	370	0.869	2.57	0.756	6.596	2.23	2.556	0.00016	359
8	8.2	500	0.914	2.69	0.835	7.284	2.47	2.693	0.00004	493
9	8.5	610	0.929	2.78	0.864	7.758	2.59	2.741	0.00201	550
Σ	56.8	2490	7.043	20.6	5.66	48.64	16.6		0.08239	

Далее рассчитаем суммы, необходимые промежуточные значения и коэффициенты (расчеты выполнялись в среде Excel):

$\Sigma Y = \Sigma \lg y = 20.6$, $\Sigma Y^2 = \Sigma(\lg y)^2 = 48.64$, $\Sigma X = \Sigma \lg x = 7.043$,

$\Sigma X^2 = \Sigma(\lg x)^2 = 5.659$, $\Sigma XY = \Sigma(\lg x \cdot \lg y) = 16.577$,

$M_Y = \Sigma Y / n = 20.6 / 9 = 2.289$, $M_X = \Sigma X / n = 7.043 / 9 = 0.7826$,

$C_{XY} = \Sigma XY - (\Sigma X) \cdot (\Sigma Y) / n = 16.572 - 7.043 \cdot 20.602 / 9 = 0.45542$,

$C_X = \Sigma X^2 - (\Sigma X)^2 / n = 5.655 - (7.04)^2 / 9 = 0.14816$,

$C_Y = \Sigma Y^2 - (\Sigma Y)^2 / n = 48.638 - (20.601)^2 / 9 = 1.4823$,

$$S_Y = \sqrt{C_Y /(n-1)} = \sqrt{1.4823/8} = 0.4305,$$

$$S_X = \sqrt{C_X /(n-1)} = \sqrt{0.14816/8} = 0.1361,$$

$$r = C_{XY} \Big/ \sqrt{C_X \cdot C_Y} = 0.45542 \Big/ \sqrt{0.14816 \cdot 1.34823} = 0.9718,$$

$a = C_{XY} / C_X = 0.45541 / 0.14815 = 3.0739$,

$B = M_Y - aM_X = 2.289 - 3.0739 \cdot 0.7826 = -0.11643$.

Линейное уравнение для преобразованных данных имеет вид:

$\lg y = 3.07 \cdot \lg x + \lg(-0.116)$ или $Y' = 3.07 \cdot X - 0.116$.

Это уравнение дает возможность рассчитать теоретические значения признака Y' (теоретические значения логарифмов массы печени), квадраты отклонений прогнозных значений от реальных: $(Y' - Y)^2$, а также их сумму $\Sigma(Y' - Y)^2 = 0.08239$.

Эта величина есть остаточная сумма квадратов; вместе с общей суммой квадратов $Cy = C_{общ.} = 1.4823$ она позволяет сформировать таблицу дисперсионного анализа (табл. 20): $C_{мод.} = C_{общ.} - C_{остат.} = 1.4823 - 0.08239 = 1.39993$.

Таблица 20

Составляющие дисперсии	C		df	S^2	F
Наклон модельной линии	$C_{регр.} = \Sigma (Y'_i - M_Y)^2$	1.399	1	$S^2_{регр.} =$ 0.39993	$F =$ 118.9377
Отклонения вариант от линии	$C_{остат.} = \Sigma (y_i - Y'_i)^2$	0.0824	6	$S^2_{остат.} =$ 0.01177	$F_{(0.05,1,7)} =$ 5.6
Общая (всего)	$C_{общ.} = \Sigma (y_i - M_Y)^2$	1.482			

Полученное значение $F = 118$ больше табличного (5.6), следовательно, дисперсия, обусловленная регрессией, достоверно больше случайной, т. е. признак Y действительно зависит от признака X, и линия регрессии адекватна исходным данным. Коэффициент детерминации больше, чем у линейной регрессии, и составляет: $R^2 = C_{регр.} / C_{общ.} = 1.399 / 1.4823 = 0.944$.

Ошибка коэффициента криволинейной регрессии равна:

$$m_a = \frac{S_y}{S_x} \cdot \sqrt{\frac{1-r^2}{n-2}} = \frac{0.430}{0.136} \cdot \sqrt{\frac{1-0.9718^2}{9-2}} = 0.281,$$

а критерий Стьюдента, проверяющий гипотезу Но: $a = 0$, составляет

$t = a / m_a = 3.0739 / 0.281 = 10.9$.

Полученное значение (10.9) больше табличного ($t_{(0.05,\,8)} = 2.31$ для уровня значимости $\alpha = 0.05$ и числа степеней свободы $df = n - 2 = 8$), коэффициент регрессии a значимо отличается от нуля; зависимость признака Y от X есть, причем очень тесная. Следует помнить, что при расчете ошибки коэффициента криволинейной регрессии используются стандартные отклонения для преобразованных (у нас – прологарифмированных) значений признаков.

В завершение выполним обратное преобразование второго коэффициента регрессии, свободный член равен:

$b = 10^B = 10^{-0.11643} = 0.764839$.

Теперь уравнение регрессии принимает вид степенной зависимости:

$y' = 0.765 \cdot x^{3.07}$.

Теоретические значения y', рассчитанные по этому уравнению, гораздо ближе к исходным данным, что хорошо видно и на графике (рис. 16), и по большей величине коэффициента детерминации ($0.94 > 0.87$) (читателю несложно будет проделать все вычисления в среде Excel с помощью программы Регрессия – как для исходных, так и для преобразованных данных).

Аллометрическое уравнение ($y' = 0.8x^3$) не только лучше описывает зависимость между признаками в статистическом плане, но и придает ей более ясный биологический смысл (масса печени = 0.8·длина тела^3). Как известно, объемные величины (объем, масса тела) пропорциональны кубу линейных промеров (длина тела). В свою очередь, вес печени и вес тела связаны прямой пропорциональной зависимостью. Так становится понятной наблюдаемая прямая пропорциональность веса печени кубу длины тела.

ВМЕСТО ПОСЛЕСЛОВИЯ

В конце книги авторы посчитали полезным поместить практически неизвестное широкому читателю стихотворное произведение «Гайавата ставит эксперимент», принадлежащее перу одного из самых крупных мировых авторитетов в области математической статистики – Мориса Дж. Кендалла. На первый взгляд оно достаточно далеко от традиционного жанра научной публикации, но на самом деле не только остроумно, иронично и талантливо само по себе, но и весьма точно отражает, разъясняет и иллюстрирует суть наиболее актуальных и сложных дискуссионных проблем современной вариационной статистики. Уж лучше Великого Кендалла обо всем этом, разумеется, не скажешь!

Предлагаемая вниманию читателей поэма впервые увидела свет на страницах журнала «American statistic» (1959, № 13) и воспроизводится в поэтическом переводе А. Дмоховского (см. Ричард Беллман «Процессы регулирования с адаптацией»: Пер. с англ. М.: Наука, 1964).

ГАЙАВАТА СТАВИТ ЭКСПЕРИМЕНТ

1

Всюду славен Гайавата,
Он стрелок непревзойденный.
Легкий лук он поднимает –
Десять стрел взмывают к небу,
И последняя слетает
С тетивы тугой, звенящей
Прежде, чем вонзится в землю
Первая из десяти.
Все, кто видел Гайавату,
Говорили, что бесспорно
Совершенства он достиг.

2

Но какой-то хитрый скептик
Тем не менее заметил,
Что в стрельбе не только ловкость,
Но и меткость ценят люди.
И добавил: было б лучше,
Если б славный Гайавата
В цель попал бы хоть однажды,
Пусть хоть выборка при этом
Будет меньшего объема.

3

Гайавата рассердился
И сказал, что он в колледже
Посвятил себя науке,
Что статистикой зовется,
Он себя считая вправе
Поучать своих собратьев,
Тут же лекцию прочел им.
Вспомнил он закон ошибок,
Усеченные кривые,
Информации потерю,
Заявил, что он добился
Несмещенных результатов,
И сказал, что после многих
Независимых попыток,
Даже если в их итоге
В цель ни разу не попал он, –
Все равно по средней точке
Отклонений от мишени
Можно сделать твердый вывод,
Что стрелял он безупречно
(За возможным исключеньем
Пресловутой меры нуль).

4

Но упрямые индейцы
Возразили Гайавате,
Что они не понимают
Столь туманных рассуждений.

Им совсем не интересен
Результат его попыток.
И они предполагают,
Что охотник должен метко
В цель стрелять. А если будет
Он впустую тратить стрелы –
Должен сам за них платить.

5

Раздраженный Гайавата
Стал цитировать обильно
Р. А. Фишера и Итса,
Приводить работы Финни,
Книги Кемпторна Оскара,
Главы Кокрана и Кокса,
Андерсена и Банкрофта.
Он взывал к авторитетам,
Убеждая несогласных,
Что в стрельбе всего важнее
Не прямое попаданье,
А научно безупречный
Статистический подход.

6

Кое-кто из возражавших
Согласился с Гайаватой,
Что в подобной точке зренья
Есть, возможно, доля смысла,
Но, пожалуй, все же лучше
Не пускаться в рассужденья,
А без промаха стрелять.

7

Наш герой в ответ на это
Предложил за луки взяться,
Чтоб строптивых оппонентов
В правоте своей уверить.
Он сказал: «Необходимо
Так построить состязанье,
Как советует учебник
Проводить эксперименты».
(Хоть научный этот способ
Применяется обычно
Для проверки качеств чая,
Но порою, как известно,
Приложим к другим вещам.)
Гайавата разработал
Точный план соревнований,
Чтоб случайный их порядок
В соответствие пришелся
С тем характером, который
Носят множители в славной
Той теории, что ныне
Носит имя Галуа.

8

Те, кто выразил готовность
Состязаться с Гайаватой,
Были круглые невежды
В проведеньи испытаний,
И поэтому, наверно,
Все оставшееся время
Проводили в тренировках,
Соревнуясь меж собою
Или просто в цель стреляя.

9

И во время состязанья
Результаты всех стрелявших
Были просто превосходны,
Но, увы, за исключеньем
(Как ни трудно мне признаться)
Результата Гайаваты.
Гайавата, как обычно,
Вверх свои направил стрелы.
Он так ловко это сделал,
Что остался несмещенным,
Но при этом, к сожаленью,
В цель ни разу не попал.

10

«Что ж, – сказали тут индейцы, –
Мы иного и не ждали».

11

Гайавата, не смущаясь,
Попросил перо, бумагу,
Произвел расчет дисперсий
И в итоге вывел цифры,
Из которых стало ясно,
Что стрелки смогли добиться
Лишь смещенных результатов,
И дисперсии при этом
Одинаковыми были
И совсем не отличались
От дисперсии, которой
Гайавата сам достиг.

(Правда, следует отметить,
Что последний этот вывод
Убедительнее был бы,
Если б в данных Гайаваты,
По которым вычислял он
Результат эксперимента,
Зафиксированы были
И прямые попаданья.
К сожаленью, оппоненты,
В вычислениях не смысля,
Не смогли с героем спорить,
Что бывает очень часто
При анализе дисперсий.)

12

Тем не менее индейцы,
Не поверившие цифрам,
Отобрали у героя
Легкий лук его и стрелы
И сказали, что, возможно,
Гайавата в самом деле
Выдающийся статистик,
Но при этом совершенно
Бесполезен как стрелок.
Что ж касается дисперсий,
То какой-то грубый неуч
Произнес такое слово,
Что его, сказать по чести,
В статистическом изданьи
Я не смею повторить.

13

И теперь в лесу дремучем
Бродит грустный Гайавата.
Непрестанно размышляя,
Вспоминает он нормальный
Тот закон распределенья
Отклонений и ошибок,
Что лишил его навеки
Славы лучшего стрелка.
И порою он приходит
К трезвой мысли, что наверно
Нужно целиться точнее,
Несмотря на риск смещенья,
Если все же в результате
Иногда ему удастся
Поражать стрелою цель.

СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ

Адлер Ю. П., Макарова Е. В., Грановский Ю. В. Планирование эксперимента при поиске оптимальных условий. М.: Наука, 1976.

Ашмарин И. П. и др. Быстрые методы статистической обработки и планирования экспериментов. Л.: Изд-во ЛГУ, 1975.

Бейли Н. Статистические методы в биологии. М.: Мир, 1964.

Браунли К. А. Статистическая теория и методология в науке и технике. М.: Наука, 1977.

Гроссман С., Терней Дж. Математика для биологов. М.: Высшая школа, 1983.

Гублер Е. В., Генкина А. А. Применение непараметрических критериев статистики в медико-биологических исследованиях. Л.: Медицина, 1973.

Дэвис Дж. Статистический анализ данных в геологии: В 2 кн. М.: Недра, 1990.

Животовский Л. А. Популяционная биометрия. М.: Наука, 1991.

Зайцев Г. Н. Математический анализ биологических данных. М.: Наука, 1981.

Зайцев Г. Н. Математика в экспериментальной ботанике. М.: Наука, 1990.

Ивантер Э. В., Коросов А. В. Введение в количественную биологию. Петрозаводск, 2003.

Коросов А. В. Экологические приложения компонентного анализа. Петрозаводск, 1996.

Лакин Г. Ф. Биометрия. М.: Высшая школа, 1973.

Плохинский Н. А. Биометрия. М.: Изд-во МГУ, 1970.

Поллард Дж. Справочник по вычислительным методам статистики. М.: Финансы и статистика, 1982.

Рокицкий П. Ф. Биологическая статистика. Минск: Вышейшая школа, 1973.

Тюрин Ю. Н., Макаров А. А. Статистический анализ данных на компьютере. М.: ИНФРА, 1998.

Урбах В. Ю. Биометрические методы. М.: Наука, 1964.

Урбах В. Ю. Статистический анализ в биологических и медицинских исследованиях. М.: Медицина, 1975.

Фишер Р. Статистические методы для исследователей. М.: Госстатиздат, 1958.

СПРАВОЧНЫЕ ТАБЛИЦЫ

Таблица 1*П*

Квадраты и квадратные корни для чисел 1…99

x	x^2	$\sqrt{x}$	x	x^2	$\sqrt{x}$	x	x^2	$\sqrt{x}$
1	1	1.000	34	1156	5.831	67	4489	8.185
2	4	1.414	35	1225	5.916	68	4624	8.246
3	9	1.732	36	1296	6.000	69	4761	8.307
4	16	2.000	37	1369	6.083	70	4900	8.367
5	25	2.236	38	1444	6.164	71	5041	8.426
6	36	2.449	39	1521	6.245	72	5184	8.485
7	49	2.646	40	1600	6.325	73	5329	8.544
8	64	2.828	41	1681	6.403	74	5476	8.602
9	81	3.000	42	1764	6.481	75	5625	8.660
10	100	3.162	43	1849	6.557	76	5776	8.718
11	121	3.317	44	1936	6.433	77	5929	8.775
12	144	3.464	45	2025	6.708	78	6084	8.832
13	169	3.606	46	2116	6.782	79	6241	8.888
14	196	3.742	47	2209	6.856	80	6400	8.944
15	225	3.873	48	2304	6.928	81	6561	9.000
16	256	4.000	49	2401	7.000	82	6724	9.055
17	289	4.123	50	2500	7.071	83	6889	9.110
18	324	4.243	51	2601	7.141	84	7056	9.165
19	361	4.359	52	2704	7.211	85	7225	9.220
20	400	4.472	53	2809	7.280	86	7396	9.274
21	441	4.583	54	2916	7.348	87	7569	9.327
22	484	4.690	55	3025	7.416	88	7744	9.381
23	529	4.796	56	3136	7.483	89	7921	9.434
24	576	4.899	57	3249	7.550	90	8100	9.487
25	625	5.000	58	3364	7.616	91	8281	9.539
26	676	5.099	59	3481	7.681	92	8464	9.592
27	729	5.196	60	3600	7.746	93	8649	9.644
28	784	5.292	61	3721	7.810	94	8836	9.695
29	841	5.385	62	3844	7.874	95	9025	9.747
30	900	5.477	63	3969	7.937	96	9216	9.798
31	961	5.568	64	4096	8.000	97	9409	9.849
32	1024	5.657	65	4225	8.062	98	9604	9.899
33	1089	5.745	66	4356	8.124	99	9801	9.950

Таблица 2П

Перевод календарных дат в непрерывный ряд

Месяцы											
III	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**XI**	**XII**	**I**	**II**
1	32	62	93	123	154	185	215	246	276	307	338
2	33	63	94	124	155	186	216	247	277	308	339
3	34	64	95	125	156	187	217	248	278	309	340
4	35	65	96	126	157	188	218	249	279	310	341
5	36	66	97	127	158	189	219	250	280	311	342
6	37	67	98	128	159	190	220	251	281	312	343
7	38	68	99	129	160	191	221	252	282	313	344
8	39	69	100	130	161	192	222	253	283	314	345
9	40	70	101	131	162	193	223	254	284	315	346
10	41	71	102	132	163	194	224	255	285	316	347
11	42	72	103	133	164	195	225	256	286	317	348
12	43	73	104	134	165	196	226	257	287	318	349
13	44	74	105	135	166	197	227	258	288	319	350
14	45	75	106	136	167	198	228	259	289	320	351
15	46	76	107	137	168	199	229	260	290	321	352
16	47	77	108	138	169	200	230	261	291	322	353
17	48	78	109	139	170	201	231	262	292	323	354
18	49	79	110	140	171	202	232	263	293	324	355
19	50	80	111	141	172	203	233	264	294	325	356
20	51	81	112	142	173	203	234	265	295	326	357
21	52	82	113	143	174	205	235	266	296	327	358
22	53	83	114	144	175	206	236	267	297	328	359
23	54	84	115	145	176	207	237	268	298	329	360
24	55	85	116	146	177	208	238	269	299	330	361
25	56	86	117	147	178	209	239	270	300	331	362
26	57	87	118	148	179	210	240	271	301	332	363
27	58	88	119	149	180	211	241	272	302	333	364
28	59	89	120	150	181	212	242	273	303	334	365
29	60	90	121	151	182	213	243	274	304	335	(366)
30	61	91	122	152	183	214	244	275	305	336	
31		92		153	184		245		306	337	

Таблица 3П

Значения случайных чисел, равномерно распределенных на интервале (0, 1)

10097	32533	76520	13586	34673	64876
37542	04865	64894	74296	24805	24037
08422	68953	19645	09303	23209	02560
99019	02529	09376	70715	38311	31165
12807	99970	80157	36147	64032	36653
80969	09117	39292	74945	66065	74717
20636	10402	00822	91665	31060	10805
15953	34764	35080	33606	85269	77602
88676	74397	04436	27659	63573	32135
98951	16877	19171	78833	73796	45753
34072	76850	36697	36170	65813	39885
45571	82406	35303	42614	86779	07439
02051	65692	68665	74818	73053	85247
05325	47048	90553	57548	28468	28709
03529	64778	35808	34282	60935	20344
11199	29170	98520	17767	14905	68607
23403	09732	11805	05431	39808	27732
18623	88579	83452	99634	06288	98083
83491	25624	88685	40200	86507	58401
35273	88435	99594	67348	87517	64960
52109	40555	60970	93433	50500	73998
50725	68248	29405	24201	52775	67851
13746	70078	18475	40610	68711	77817
36766	67951	90364	76493	29609	11062
91826	08928	93785	61368	23478	34113
65481	17674	17468	50950	79335	51748
80124	35635	17727	08015	82391	90324

Таблица 4П

Ординаты нормальной кривой

(значения функции $f(t)=\frac{1}{\sqrt{2\pi}}\cdot e^{\frac{t^2}{2}}$ **)**

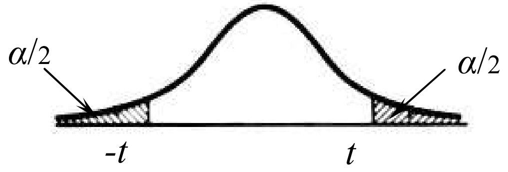
t	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.3989	0.3989	0.3989	0.3988	0.3986	0.3984	0.3982	0.3980	0.3977	0.3973
0.1	0.3970	0.3965	0.3961	0.3956	0.3951	0.3945	0.3939	0.3932	0.3825	0.3918
0.2	0.3910	0.3902	0.3894	0.3885	0.3876	0.3867	0.3857	0.3847	0.3836	0.3825
0.3	0.3814	0.3802	0.3790	0.3778	0.3765	0.3752	0.3739	0.3726	0.3712	0.3697
0.4	0.3683	0.3668	0.3653	0.3637	0.3621	0.3605	0.3589	0.3572	0.3555	0.3538
0.5	0.3521	0.3503	0.3485	0.3467	0.3448	0.3429	0.3410	0.3391	0.3372	0.3352
0.6	0.3332	0.3312	0.3292	0.3271	0.3251	0.3230	0.3209	0.3187	0.3166	0.3144
0.7	0.3123	0.3101	0.3079	0.3056	0.3034	0.3011	0.2989	0.2966	0.2943	0.2920
0.8	0.2987	0.2874	0.2850	0.2827	0.2803	0.2780	0.2756	0.2732	0.2709	0.2685
0.9	0.2661	0.2637	0.2613	0.2589	0.2565	0.2541	0.2516	0.2492	0.2468	0.2444
1	0.2420	0.2396	0.2371	0.2347	0.2323	0.2299	0.2275	0.2251	0.2227	0.2203
1.1	0.2179	0.2155	0.2131	0.2107	0.2083	0.2059	0.2036	0.2012	0.1989	0.1965
1.2	0.1942	0.1919	0.1895	0.1872	0.1849	0.1826	0.1804	0.1781	0.1758	0.1736
1.3	0.1714	0.1691	0.1669	0.1647	0.1626	0.1604	0.1582	0.1561	0.1539	0.1518
1.4	0.1497	0.1476	0.1456	0.1435	0.1415	0.1394	0.1374	0.1354	0.1334	0.1315
1.5	0.1295	0.1276	0.1257	0.1238	0.1219	0.1200	0.1182	0.1163	0.1145	0.1127
1.6	0.1109	0.1092	0.1074	0.1057	0.1040	0.1023	0.1006	0.0989	0.0973	0.0957
1.7	0.0940	0.0925	0.0909	0.0893	0.0878	0.0863	0.0848	0.0833	0.0818	0.0804
1.8	0.0790	0.0775	0.0761	0.0748	0.0734	0.0721	0.0707	0.0694	0.0681	0.0669
1.9	0.0656	0.0644	0.0632	0.0620	0.0608	0.0596	0.0584	0.0573	0.0562	0.0551
2	0.0540	0.0529	0.0519	0.0508	0.0498	0.0488	0.0478	0.0468	0.0459	0.0449
2.1	0.0440	0.0431	0.0422	0.0413	0.0404	0.0396	0.0387	0.0379	0.0371	0.0363
2.2	0.0355	0.0347	0.0339	0.0332	0.0325	0.0317	0.0310	0.0303	0.0297	0.0290
2.3	0.0283	0.0277	0.0270	0.0264	0.0258	0.0252	0.0246	0.0241	0.0235	0.0229
2.4	0.0224	0.0219	0.0213	0.0208	0.0203	0.0198	0.0191	0.0189	0.0184	0.0180
2.5	0.0175	0.0171	0.0167	0.0163	0.0158	0.0154	0.0151	0.0147	0.0143	0.0139
2.6	0.0136	0.0132	0.0129	0.0126	0.0122	0.0119	0.0116	0.0113	0.0110	0.0107
2.7	0.0104	0.0101	0.0099	0.0096	0.0093	0.0091	0.0088	0.0086	0.0084	0.0081
2.8	0.0079	0.0077	0.0075	0.0073	0.0071	0.0069	0.0067	0.0065	0.0063	0.0061
2.9	0.0060	0.0058	0.0056	0.0055	0.0053	0.0051	0.0050	0.0048	0.0047	0.0046
3	0.0044	0.0043	0.0042	0.0041	0.0039	0.0038	0.0037	0.0036	0.0035	0.0034
3.1	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026	0.0025	0.0025
3.2	0.0024	0.0023	0.0022	0.0022	0.0021	0.0020	0.0020	0.0019	0.0018	0.0018
3.3	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014	0.0013	0.0013
3.4	0.0012	0.0012	0.0012	0.0011	0.0011	0.0010	0.0010	0.0010	0.0009	0.0009
3.5	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007	0.0007	0.0007	0.0006

Таблица 5П

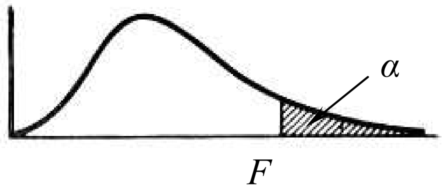
Значение критерия *t* для отбраковки «выскакивающих» вариант

n	α			*n*	α		
	0.05	**0.01**	**0.001**		**0.05**	**0.01**	**0.001**
5	3.04	5.04	9.43	20	2.15	2.93	3.98
6	2.78	4.36	7.41	25	2.11	2.85	3.82
7	2.62	3.96	6.37	30	2.08	2.80	3.72
8	2.51	3.71	5.73	35	2.06	2.77	3.65
9	2.43	3.54	5.31	40	2.05	2.74	3.60
10	2.37	3.41	5.01	45	2.04	2.72	3.57
11	2.33	3.31	4.79	50	2.03	2.71	3.53
12	2.29	3.23	4.62	60	2.02	2.68	3.49
13	2.26	3.17	4.48	70	2.01	2.67	3.46
14	2.24	3.12	4.37	80	2.00	2.66	3.44
15	2.22	3.08	4.28	90	2.00	2.65	3.42
16	2.20	3.04	4.20	100	1.99	2.64	3.41
17	2.18	3.01	4.13	0	1.96	2.58	3.29
18	2.17	2.98	4.07				

Пороговые значения распределения *T* Стьюдента;
α для двустороннего критерия



Пороговые значения распределения *F* Фишера



Пороговые значения распределения χ^2 Пирсона

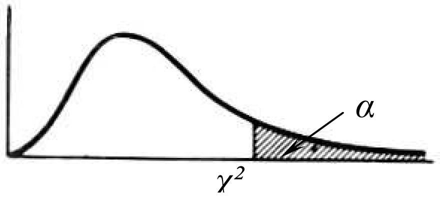


Таблица 6П

Значения критерия *t* Стьюдента

Число степеней свободы, *df*	Доверительная вероятность (*P*) Уровень значимости (α)		
	P = 0.095 $\alpha = 0.05$	$P = 0.099$ $\alpha = 0.01$	$P = 0.0999$ $\alpha = 0.001$
2	4.303	9.925	31.598
3	3.182	5.841	12.941
4	2.776	4.604	8.610
5	2.571	4.032	6.859
6	2.447	3.707	5.959
7	2.365	3.499	5.405
8	2.306	3.355	5.041
9	2.262	3.250	4.781
10	2.228	3.169	4.587
11	2.201	3.106	4.437
12	2.179	3.055	4.318
13	2.160	3.012	4.221
14	2.145	2.977	4.140
15	2.131	2.947	4.073
16	2.120	2.921	4.015
17	2.110	2.898	3.965
18	2.101	2.878	3.922
19	2.093	2.861	3.883
20	2.086	2.845	3.850
22	2.074	2.819	3.792
25	2.060	2.787	3.725
30	2.042	2.750	3.646
35	2.030	2.724	3.591
40	2.021	2.704	3.551
45	2.014	2.690	3.520
50	2.008	2.678	3.496
55	2.004	2.669	3.476
60	2.000	2.660	3.460
70	1.994	2.648	3.435
80	1.989	2.638	3.416
90	1.986	2.631	3.402
100	1.982	2.625	3.390
120	1.980	2.617	3.373
>120	1.960	2.5758	3.2905

Таблица 7П

Значения критерия *F* Фишера при уровне значимости α = 0.05 (число степеней свободы указано для дисперсии знаменателя – в строке, для дисперсии числителя – в столбце)

df_1 / df_2	1	2	3	4	5	6	7	8	9	10	15	20	30	∞
1	161	200	216	225	230	234	237	239	241	242	246	248	250	254
2	18.5	19.0	19.2	19.3	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.4
3	10.1	9.6	9.3	9.1	9.0	8.9	8.9	8.9	8.8	8.8	8.7	8.7	8.6	8.5
4	7.7	6.9	6.6	6.4	6.3	6.2	6.1	6.0	6.0	5.9	5.9	5.8	5.8	5.6
5	6.6	5.8	5.4	5.2	5.1	5.0	4.9	4.8	4.8	4.7	4.6	4.6	4.5	4.4
6	6.0	5.1	4.7	4.5	4.4	4.3	4.2	4.2	4.1	4.1	4.0	3.9	3.8	3.7
7	5.6	4.7	4.4	4.1	4.0	3.9	3.8	3.7	3.7	3.6	3.5	3.4	3.4	3.2
8	5.3	4.5	4.1	3.8	3.7	3.6	3.5	3.4	3.4	3.3	3.2	3.2	3.1	3.0
9	5.1	4.3	3.9	3.6	3.5	3.4	3.3	3.2	3.2	3.1	3.0	2.9	2.9	2.7
10	5.0	4.1	3.7	3.5	3.3	3.2	3.1	3.1	3.0	3.0	2.9	2.8	2.7	2.5
11	4.8	4.0	3.6	3.4	3.2	3.1	3.0	3.0	2.9	2.9	2.7	2.7	2.6	2.4
12	4.7	3.9	3.5	3.3	3.1	3.0	2.9	2.9	2.8	2.8	2.6	2.5	2.5	2.3
13	4.7	3.8	3.4	3.2	3.0	2.9	2.8	2.8	2.7	2.7	2.5	2.5	2.4	2.2
14	4.6	3.7	3.3	3.1	3.0	2.9	2.8	2.7	2.7	2.6	2.5	2.4	2.3	2.1
15	4.5	3.7	3.3	3.1	2.9	2.8	2.7	2.6	2.6	2.5	2.4	2.3	2.2	2.1
16	4.5	3.6	3.2	3.0	2.8	2.7	2.7	2.6	2.5	2.5	2.3	2.3	2.2	2.0
17	4.4	3.6	3.2	2.9	2.8	2.7	2.6	2.5	2.5	2.4	2.3	2.2	2.1	2.0
18	4.4	3.5	3.2	2.9	2.8	2.7	2.6	2.5	2.5	2.4	2.3	2.2	2.1	1.9
19	4.4	3.5	3.1	2.9	2.7	2.6	2.5	2.5	2.4	2.4	2.2	2.2	2.1	1.9
20	4.3	3.5	3.1	2.9	2.7	2.6	2.5	2.4	2.4	2.3	2.2	2.1	2.0	1.8
21	4.3	3.5	3.1	2.8	2.7	2.6	2.5	2.4	2.4	2.3	2.2	2.1	2.0	1.8
22	4.3	3.4	3.0	2.8	2.7	2.5	2.5	2.4	2.3	2.3	2.1	2.1	2.0	1.8
23	4.3	3.4	3.0	2.8	2.6	2.5	2.4	2.4	2.3	2.3	2.1	2.0	1.9	1.8
24	4.3	3.4	3.0	2.8	2.6	2.5	2.4	2.4	2.3	2.2	2.1	2.0	1.9	1.7
26	4.2	3.4	3.0	2.7	2.6	2.5	2.4	2.3	2.3	2.2	2.1	2.0	1.9	1.7
28	4.2	3.3	2.9	2.7	2.6	2.4	2.4	2.3	2.2	2.2	2.0	2.0	1.9	1.6
30	4.2	3.3	2.9	2.7	2.5	2.4	2.3	2.3	2.2	2.2	2.0	1.9	1.8	1.6
40	4.1	3.2	2.8	2.6	2.4	2.3	2.2	2.2	2.1	2.1	1.9	1.8	1.7	1.5
60	4.0	3.1	2.8	2.5	2.4	2.2	2.2	2.1	2.0	2.0	1.8	1.7	1.6	1.4
120	3.9	3.1	2.7	2.4	2.3	2.2	2.1	2.0	2.0	1.9	1.7	1.7	1.6	1.2
∞	3.8	3.0	2.6	2.4	2.2	2.1	2.0	1.9	1.9	1.8	1.7	1.6	1.5	1.0

Таблица 8П

Значения критерия *F* Фишера при уровне значимости $\alpha = 0.01$ (число степеней свободы указано для дисперсии знаменателя – в строке, для дисперсии числителя – в столбце)

df_2 \ df_1	1	2	3	4	5	6	7	8	9	10	15	20	30	∞
1	4052	4999	5403	5625	5764	5859	5928	5982	6022	6056	6157	6209	6261	6366
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5
3	31.4	30.8	29.5	28.7	28.4	27.9	27.7	27.5	27.3	27.2	26.9	26.7	26.5	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.2	14.0	13.8	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.0	9.7	9.5	9.4	9.0
6	13.7	10.9	9.8	9.1	8.7	8.5	8.3	8.1	8.0	7.9	7.6	7.4	7.2	6.9
7	12.3	9.5	8.5	7.8	7.5	7.2	7.0	6.8	6.7	6.6	6.3	6.2	6.0	5.6
8	11.3	8.7	7.6	7.0	6.6	6.4	6.2	6.0	5.9	5.8	5.5	5.4	*5.2*	4.9
9	10.6	8.0	7.0	6.4	6.1	5.8	5.6	5.5	5.3	5.3	5.0	4.8	4.6	4.3
10	10.0	7.6	6.5	6.0	5.6	5.4	5.2	5.1	4.9	4.8	4.6	4.4	4.2	3.9
11	9.7	7.2	6.2	5.7	5.3	5.1	4.9	4.7	4.6	4.5	4.2	4.1	3.9	3.6
12	9.3	6.9	5.9	5.4	5.1	4.8	4.6	4.5	4.4	4.3	4.0	3.9	3.7	3.4
13	9.1	6.7	5.7	5.2	4.9	4.6	4.4	4.3	4.2	4.1	3.8	3.7	3.5	3.2
14	8.9	6.5	5.6	5.0	4.7	4.5	4.3	4.1	4.0	3.9	3.7	3.5	3.3	3.0
15	8.7	6.4	5.4	4.9	4.6	4.3	4.1	4.0	3.9	3.8	3.5	3.4	3.2	2.9
16	8.5	6.2	5.3	4.8	4.4	4.2	4.0	3.9	3.8	3.7	3.4	3.3	3.1	2.7
17	8.4	6.1	5.2	4.7	4.3	4.1	3.9	3.8	3.7	3.6	3.3	3.2	3.0	2.6
18	8.3	6.0	5.1	4.6	4.2	4.0	3.8	3.7	3.6	3.5	3.2	3.1	2.9	2.6
19	8.2	5.9	5.0	4.5	4.2	3.9	3.8	3.6	3.5	3.4	3.1	3.0	2.8	2.5
20	8.1	5.8	4.9	4.4	4.1	3.9	3.7	3.6	3.5	3.4	3.1	2.9	2.8	2.4
21	8.0	5.8	4.9	4.4	4.0	3.8	3.6	3.5	3.4	3.3	3.0	2.9	2.7	2.4
22	7.9	5.7	4.8	4.3	4.0	3.8	3.6	3.4	3.3	3.3	3.0	2.8	2.7	2.3
23	7.9	5.7	4.8	4.3	3.9	3.7	3.5	3.4	3.3	3.2	2.9	2.7	2.6	2.3
24	7.8	5.6	4.7	4.2	3.9	3.7	3.5	3.4	3.3	3.2	2.9	2.7	2.6	2.2
26	7.7	5.5	4.6	4.1	3.8	3.6	3.4	3.3	3.2	3.1	2.8	2.7	2.5	2.1
28	7.6	5.4	4.6	4.1	3.7	3.5	3.4	3.2	3.1	3.0	2.7	2.6	2.4	2.1
30	7.6	5.4	4.5	4.0	3.7	3.5	3.3	3.2	3.1	3.0	2.7	2.5	2.4	2.0
40	7.3	5.2	4.3	3.8	3.5	3.3	3.1	3.0	2.9	2.8	2.5	2.4	2.2	1.8
60	7.1	5.0	4.1	3.6	3.3	3.1	2.9	2.8	2.7	2.6	2.3	2.2	2.0	1.6
120	6.8	4.8	3.9	3.5	3.2	3.0	2.8	2.7	2.6	2.5	2.2	2.0	1.9	1.4
∞	6.6	4.6	3.8	3.3	3.0	2.8	2.6	2.5	2.4	2.3	2.5	1.9	1.7	1.0

Таблица 9П

Значения критерия χ^2 Пирсона

df	Уровень значимости, *α*				
	0.95	**0.75**	**0.25**	***0.05***	**0.01**
1	–	0.10	1.32	3.84	6.63
2	0.10	0.58	2.77	5.99	9.21
3	0.35	1.21	4.11	7.81	11.34
4	0.71	1.92	5.39	9.49	13.28
5	1.15	2.67	6.63	11.07	15.09
6	1.64	3.45	7.84	12.59	16.81
7	2.17	4.25	9.04	14.07	18.48
8	2.73	5.07	10.22	15.51	20.09
9	3.33	5.90	11.39	16.92	21.67
10	3.94	6.74	12.55	18.31	23.21
11	4.57	7.58	13.70	19.68	24.72
12	5.23	8.44	14.85	21.03	26.22
13	5.89	9.30	15.98	22.36	27.69
14	6.57	10.17	17.12	23.68	29.14
15	7.26	11.04	18.25	25.00	30.58
16	7.96	11.91	19.37	26.30	32.00
17	8.67	12.79	20.49	27.59	33.41
18	9.39	13.68	21.60	28.87	34.81
19	10.12	14.56	22.72	30.14	36.19
20	10.85	15.45	23.83	31.41	37.57
21	11.59	16.34	24.93	32.67	38.93
22	12.34	17.24	26.04	33.92	40.29
23	13.09	18.14	27.14	35.17	41.64
24	13.85	19.04	28.24	36.42	42.98
25	14.61	19.94	29.34	37.65	44.31
26	15.38	20.84	30.43	38.89	45.64
27	16.15	21.75	31.63	40.11	46.96
28	16.93	22.66	32.62	41.34	48.28
30	18.49	24.48	34.80	43.77	50.89
40	26.51	33.66	45.62	55.76	63.69
50	34.76	42.94	56.33	67.50	76.15
60	43.19	52.29	66.98	79.08	88.38
70	51.74	61.70	77.58	90.53	100.42
80	60.39	71.14	88.13	101.88	112.33
90	69.13	80.62	98.64	113.14	124.12
100	77.93	90.13	109.14	124.34	135.81

Таблица 10*П*

Значения $\varphi = 2\arcsin\sqrt{p}$

p, %	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.000	0.063	0.089	0.110	0.127	0.142	0.155	0.168	0.179	0.190
1	0.200	0.210	0.220	0.229	0.237	0.246	0.254	0.262	0.269	0.277
2	0.284	0.291	0.298	0.304	0.311	0.318	0.324	0.330	0.336	0.342
3	0.348	0.354	0.360	0.363	0.371	0.376	0.382	0.387	0.392	0.398
4	0.403	0.408	0.413	0.418	0.423	0.428	0.432	0.437	0.442	0.448
5	0.451	0.456	0.460	0.465	0.469	0.473	0.478	0.482	0.486	0.491
6	0.495	0.499	0.503	0.507	0.512	0.516	0.520	0.524	0.528	0.532
7	0.536	0.539	0.543	0.546	0.551	0.555	0.559	0.562	0.566	0.570
8	0.574	0.577	0.581	0.584	0.588	0.592	0.595	0.599	0.602	0.606
9	0.609	0.613	0.616	0.620	0.623	0.627	0.630	0.633	0.637	0.640
10	0.644	0.647	0.650	0.653	0.657	0.660	0.663	0.666	0.670	0.673
11	0.676	0.679	0.682	0.686	0.689	0.692	0.695	0.698	0.701	0.704
12	0.707	0.711	0.714	0.717	0.720	0.723	0.726	0.729	0.732	0.735
13	0.738	0.741	0.744	0.747	0.750	0.752	0.755	0.758	0.761	0.764
14	0.767	0.770	0.773	0.776	0.778	0.781	0.784	0.787	0.790	0.793
15	0.795	0.798	0.801	0.804	0.807	0.809	0.812	0.815	0.818	0.820
16	0.823	0.826	0.828	0.831	0.834	0.837	0.839	0.842	0.845	0.847
17	0.850	0.853	0.855	0.858	0.861	0.863	0.866	0.868	0.871	0.874
18	0.876	0.879	0.881	0.884	0.887	0.889	0.892	0.894	0.897	0.900
19	0.902	0.905	0.907	0.910	0.912	0.915	0.917	0.920	0.922	0.925
20	0.927	0.930	0.932	0.935	0.937	0.940	0.942	0.945	0.947	0.950
21	0.952	0.955	0.957	0.959	0.962	0.964	0.967	0.969	0.972	0.974
22	0.976	0.979	0.981	0.984	0.986	0.988	0.991	0.993	0.996	0.998
23	1.000	1.003	1.005	1.007	1.010	1.012	1.015	1.017	1.019	1.022
24	1.024	1.026	1.029	1.031	1.033	1.036	1.038	1.040	1.043	1.045
25	1.047	1.050	1.052	1.054	1.056	1.059	1.061	1.063	1.066	1.068
26	1.070	1.072	1.075	1.077	1.079	1.082	1.084	1.086	1.088	1.091
27	1.093	1.095	1.097	1.100	1.102	1.104	1.106	1.109	1.111	1.113
28	1.115	1.117	1.120	1.122	1.124	1.126	1.129	1.131	1.133	1.135
29	1.137	1.140	1.142	1.144	1.146	1.148	1.151	1.153	1.155	1.157
30	1.159	1.161	1.164	1.166	1.168	1.170	1.172	1.174	1.177	1.179
31	1.182	1.183	1.185	1.187	1.190	1.192	1.194	1.196	1.198	1.200
32	1.203	1.205	1.207	1.209	1.211	1.213	1.215	1.217	1.220	1.222
33	1.224	1.226	1.228	1.230	1.232	1.234	1.237	1.289	1.241	1.243
34	1.245	1.247	1.249	1.251	1.254	1.256	1.258	1.260	1.262	1.264
35	1.266	1.268	1.270	1.272	1.274	1.277	1.279	1.281	1.283	1.285
36	1.287	1.289	1.291	1.293	1.295	1.297	1.299	1.302	1.304	1.306
37	1.308	1.310	1.312	1.314	1.316	1.318	1.320	1.322	1.324	1.326

Продолжение табл. 10*П*

Значения $\varphi = 2\arcsin\sqrt{p}$

p, %	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
38	1.328	1.330	1.333	1.335	1.337	1.339	1.341	1.343	1.345	1.347
39	1.349	1.351	1.353	1.355	1.357	1.359	1.361	1.363	1.365	1.367
40	1.369	1.371	1.374	1.376	1.378	1.380	1.382	1.384	1.346	1.388
41	1.390	1.392	1.394	1.396	1.398	1.400	1.402	1.404	1.406	1.408
42	1.410	1.412	1.414	1.416	1.418	1.420	1.422	1.424	1.426	1.428
43	1.430	1.432	1.434	1.436	1.438	1.440	1.442	1.444	1.446	1.448
44	1.451	1.453	1.455	1.457	1.459	1.461	1.463	1.465	1.466	1.469
45	1.471	1.473	1.475	1.477	1.479	1.481	1.483	1.485	1.487	1.489
46	1.491	1.493	1.495	1.497	1.499	1.501	1.503	1.505	1.507	1.509
47	1.511	1.513	1.515	1.517	1.519	1.521	1.523	1.525	1.527	1.529
48	1.531	1.533	1.535	1.537	1.539	1.541	1.543	1.545	1.547	1.549
49	1.551	1.553	1.555	1.557	1.559	1.561	1.563	1.565	1.567	1.569
50	1.571	1.573	1.575	1.577	1.579	1.581	1.583	1.585	1.587	1.589
51	1.591	1.593	1.595	1.597	1.599	1.601	1.603	1.605	1.607	1.609
52	1.611	1.613	1.615	1.617	1.619	1.621	1.623	1.625	1.627	1.629
53	1.631	1.633	1.635	1.637	1.639	1.641	1.643	1.645	1.647	1.649
54	1.651	1.653	1.655	1.657	1.659	1.661	1.663	1.665	1.667	1.669
55	1.671	1.673	1.675	1.677	1.679	1.681	1.683	1.685	1.687	1.689
56	1.691	1.693	1.695	1.697	1.699	1.701	1.703	1.705	1.707	1.709
57	1.711	1.713	1.715	1.717	1.719	1.721	1.723	1.725	1.727	1.729
58	1.731	1.734	1.736	1.738	1.740	1.742	1.744	1.746	1.748	1.750
59	1.752	1.754	1.756	1.758	1.760	1.762	1.764	1.766	1.768	1.770
60	1.772	1.774	1.776	1.778	1.780	1.782	1.784	1.786	1.789	1.791
61	1.793	1.795	1.797	1.799	1.801	1.803	1.805	1.807	1.809	1.811
62	1.813	1.815	1.817	1.819	1.821	1.823	1.826	1.828	1.830	1.832
63	1.834	1.836	1.838	1.840	1.842	1.844	1.846	1.848	1.850	1.853
64	1.855	1.857	1.859	1.861	1.863	1.865	1.867	1.869	1.871	1.873
65	1.875	1.878	1.880	1.882	1.884	1.886	1.888	1.890	1.892	1.894
66	1.897	1.899	1.901	1.903	1.905	1.907	1.909	1.911	1.913	1.916
67	1.918	1.920	1.922	1.924	1.926	1.928	1.930	1.933	1.935	1.937
68	1.939	1.941	1.943	1.946	1.948	1.950	1.952	1.954	1.956	1.958
69	1.961	1.963	1.965	1.967	1.969	1.971	1.974	1.976	1.978	1.980
70	1.982	1.984	1.987	1.989	1.991	1.993	1.995	1.998	2.000	2.002
71	2.004	2.006	2.009	2.011	2.013	2.015	2.018	2.020	2.022	2.024
72	2.026	2.029	2.031	2.033	2.035	2.038	2.040	2.042	2.044	2.047
73	2.049	2.051	2.053	2.056	2.058	2.060	2.062	2.065	2.067	2.069
74	2.071	2.074	2.076	2.078	2.081	2.083	2.085	2.087	2.090	2.092
75	2.094	2.097	2.099	2.101	2.104	2.106	2.108	2.111	2.113	2.115

Окончание табл. 10*П*

Значения $\varphi = 2\arcsin\sqrt{p}$

p, %	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
76	2.118	2.120	2.122	2.125	2.127	2.129	2.132	2.134	2.136	2.139
77	2.141	2.144	2.146	2.148	2.151	2.153	2.156	2.158	2.160	2.163
78	2.165	2.168	2.170	2.172	2.175	2.177	2.180	2.182	2.185	2.187
79	2.190	2.192	2.194	2.197	2.199	2.202	2.204	2.207	2.209	2.212
80	2.214	2.217	2.219	2.222	2.224	2.222	2.229	2.231	2.234	2.237
81	2.240	2.242	2.245	2.247	2.250	2.262	2.255	2.258	2.260	2.263
82	2.265	2.268	2.271	2.273	2.276	2.278	2.281	2.284	2.286	2.289
83	2.292	2.294	2.297	2.300	2.302	2.305	2.308	2.310	2.313	2.316
84	2.319	2.321	2.324	2.327	2.330	2.332	2.335	2.338	2.341	2.343
85	2.346	2.349	2.352	2.355	2.357	2.360	2.363	2.366	2.369	2.372
86	2.375	2.377	2.380	2.383	2.386	2.389	2.392	2.395	2.398	2.402
87	2.404	2.407	2.410	2.413	2.416	2.419	2.422	2.425	2.428	2.431
88	2.434	2.437	2.440	2.443	2.447	2.450	2.453	2.456	2.459	2.462
89	2.465	2.469	2.472	2.475	2.478	2.482	2.485	2.488	2.491	2.495
90	2.498	2.501	2.505	2.508	2.512	2.515	2.518	2.522	2.525	2.529
91	2.532	2.536	2.539	2.543	2.546	2.550	2.554	2.557	2.561	2.564
92	2.568	2.572	2.575	2.579	2.583	2.587	2.591	2.594	2.598	2.600
93	2.606	2.610	2.614	2.618	2.622	2.626	2.630	2.634	2.638	2.642
94	2.647	2.651	2.655	2.659	2.664	2.668	2.673	2.677	2.638	2.642
95	2.691	2.695	2.700	2.705	2.709	2.714	2.719	2.724	2.729	2.734
96	2.739	2.744	2.749	2.754	2.760	2.765	2.771	2.776	2.782	2.788
97	2.793	2.799	2.805	2.811	2.818	2.824	2.830	2.837	2.844	2.851
98	2.858	2.865	2.872	2.880	2.888	2.896	2.904	2.913	2.922	2.931
99	2.941	2.952	2.963	2.974	2.987	3.000	3.015	3.032	3.052	3.078
100	3.142									

Таблица 11*П*

Значения критерия *U* Уилкоксона – Манна – Уитни

Уровень значимости $\alpha = 0.05$							
n	4	5	6	7	8	9	10
4	10	11	12	13	14	15	15
5		17	18	20	21	22	23
6			26	27	29	31	32
7				36	38	40	42
8					49	51	53
9						63	65
10							78

Уровень значимости $\alpha = 0.01$							
n	4	5	6	7	8	9	10
4			10	10	11	11	12
5		15	16	17	17	18	19
6			23	24	25	23	27
7				32	34	35	37
8					43	45	47
9						56	58
10							71

Таблица 12*П*

Значения критерия *Q* Розенбаума

n_2 \ n_1	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Уровень значимости α = 0.05																
11	6															
12	6	6														
13	6	6	6													
14	7	7	6	6												
15	7	7	6	6	6											
16	7	7	7	7	6	6										
17	7	7	7	7	7	7	7									
18	7	7	7	7	7	7	7	7								
19	7	7	7	7	7	7	7	7	7							
20	7	7	7	7	7	7	7	7	7	7						
21	8	7	7	7	7	7	7	7	7	7	7					
22	8	7	7	7	7	7	7	7	7	7	7	7				
23	8	8	7	7	7	7	7	7	7	7	7	7	7			
24	8	8	8	8	8	8	8	8	8	8	7	7	7	7		
25	8	8	8	8	8	8	8	8	8	8	7	7	7	7	7	
26	8	8	8	8	8	8	8	8	8	8	7	7	7	7	7	7
Уровень значимости α = 0.01																
11	9															
12	9	9														
13	9	9	9													
14	9	9	9	9												
15	9	9	9	9	9											
16	9	9	9	9	9	9										
17	10	9	9	9	9	9	9									
18	10	10	9	9	9	9	9	9								
19	10	10	10	9	9	9	9	9	9							
20	10	10	10	10	9	9	9	9	9	9						
21	11	10	10	10	9	9	9	9	9	9	9					
22	11	11	10	10	10	9	9	9	9	9	9	9				
23	11	11	10	10	10	10	9	9	9	9	9	9	9			
24	12	11	11	10	10	10	10	9	9	9	9	9	9	9		
25	12	11	11	10	10	10	10	10	9	9	9	9	9	9	9	
26	12	12	11	11	10	10	10	10	10	9	9	9	9	9	9	9

Таблица 13*П*

Значения величины $z = 0.5 \cdot \ln \frac{1+r}{1-r}$

r	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.0000	0.0100	0.0200	0.0300	0.0400	0.0501	0.0601	0.0701	0.0802	0.0902
0.1	0.1003	0.1105	0.1206	0.1308	0.1409	0.1511	0.1614	0.1717	0.1820	0.1923
0.2	0.2027	0.2132	0.2237	0.2342	0.2448	0.2554	0.2661	0.2769	0.2877	0.2986
0.3	0.3095	0.3206	0.3317	0.3428	0.3541	0.3654	0.3769	0.3884	0.4001	0.4118
0.4	0.4236	0.4356	0.4477	0.4599	0.4722	0.4847	0.4973	0.5101	0.5230	0.5361
0.5	0.5493	0.5627	0.5763	0.5901	0.6042	0.6184	0.6328	0.6475	0.6625	0.6777
0.6	0.6931	0.7089	0.7250	0.7414	0.7582	0.7753	0.7928	0.8107	0.8291	0.8480
0.7	0.8673	0.8872	0.9076	0.9287	0.9505	0.9730	0.9962	1.0203	1.0454	1.0714
0.8	1.0986	1.1270	1.1518	1.1881	1.2212	1.2562	1.2933	1.3331	1.3758	1.4219
0.9	1.4722	1.5275	1.5890	1.6584	1.7380	1.8318	1.9459	2.0923	2.2976	2.6467

Таблица 14*П*

Значения *r* для *z* от 0.00 до 1.99

z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.000	0.010	0.020	0.030	0.040	0.050	0.060	0.070	0.080	0.090
0.1	0.100	0.110	0.119	0.129	0.139	0.149	0.159	0.168	0.178	0.188
0.2	0.197	0.207	0.217	0.226	0.236	0.245	0.254	0.264	0.273	0.282
0.3	0.291	0.300	0.310	0.319	0.328	0.336	0.345	0.354	0.363	0.371
0.4	0.380	0.389	0.397	0.405	0.414	0.422	0.430	0.438	0.446	0.454
0.5	0.462	0.470	0.478	0.485	0.493	0.501	0.508	0.515	0.523	0.530
0.6	0.537	0.544	0.551	0.558	0.565	0.572	0.578	0.585	0.592	0.598
0.7	0.604	0.611	0.617	0.623	0.629	0.635	0.641	0.647	0.653	0.658
0.8	0.664	0.670	0.675	0.681	0.686	0.691	0.696	0.701	0.706	0.711
0.9	0.716	0.721	0.726	0.731	0.735	0.740	0.744	0.749	0.753	0.757
1.0	0.762	0.766	0.770	0.774	0.778	0.782	0.786	0.790	0.793	0.797
1.1	0.801	0.804	0.808	0.811	0.814	0.818	0.821	0.824	0.828	0.831
1.2	0.834	0.837	0.840	0.843	0.846	0.848	0.851	0.854	0.857	0.859
1.3	0.862	0.864	0.867	0.869	0.872	0.874	0.876	0.879	0.881	0.883
1.4	0.885	0.888	0.890	0.892	0.894	0.896	0.898	0.900	0.902	0.903
1.5	0.905	0.907	0.909	0.910	0.912	0.914	0.915	0.917	0.919	0.920
1.6	0.922	0.923	0.925	0.926	0.928	0.929	0.930	0.932	0.933	0.934
1.7	0.935	0.937	0.938	0.939	0.940	0.941	0.943	0.944	0.945	0.946
1.8	0.947	0.948	0.949	0.950	0.951	0.952	0.953	0.954	0.955	0.955
1.9	0.956	0.957	0.958	0.959	0.960	0.960	0.961	0.962	0.963	0.963

Окончание табл. 14*П*

Значения *r* для *z* от 2.00 до 2.99

z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
2.0	0.964	0.965	0.965	0.966	0.967	0.967	0.968	0.969	0.969	0.970
2.1	0.970	0.972	0.972	0.972	0.973	0.973	0.974	0.974	0.975	0.975
2.2	0.976	0.976	0.977	0.977	0.978	0.978	0.979	0.979	0.979	0.980
2.3	0.980	0.981	0.981	0.981	0.982	0.982	0.982	0.933	0.983	0.983
2.4	0.984	0.934	0.984	0.985	0.935	0.935	0.936	0.986	0.986	0.986
2.5	0.987	0.987	0.987	0.987	0.988	0.988	0.938	0.988	0.989	0.989
2.6	0.989	0.939	0.989	0.990	0.990	0.990	0.990	0.990	0.991	0.991
2.7	0.991	0.991	0.991	0.992	0.992	0.992	0.992	0.992	0.992	0.993
2.8	0.993	0.993	0.993	0.993	0.993	0.993	0.994	0.994	0.994	0.994

Таблица 15*П*

Минимальные значения коэффициента корреляции *r*, достоверно отличные от нуля (*df* = *n* − 2)

df	*α*		*df*	*α*		*df*	*α*	
	0.05	**0.01**		**0.05**	**0.01**		**0.05**	**0.01**
1	0.997	1	**16**	0.468	0.59	**40**	0.304	0.393
2	0.95	0.99	**17**	0.456	0.575	**45**	0.288	0.372
3	0.878	0.959	**18**	0.444	0.561	**50**	0.273	0.354
4	0.811	0.917	**19**	0.433	0.549	**60**	0.25	0.325
5	0.754	0.874	**20**	0.423	0.537	**70**	0.232	0.302
6	0.707	0.834	**21**	0.413	0.526	**80**	0.217	0.283
7	0.666	0.798	**22**	0.404	0.515	**90**	0.205	0.267
8	0.632	0.765	**23**	0.396	0.505	**100**	0.195	0.254
9	0.602	0.735	**24**	0.388	0.496	**125**	0.174	0.228
10	0.576	0.708	**25**	0.381	0.487	**150**	0.159	0.208
11	0.553	0.684	**26**	0.374	0.478	**200**	0.138	0.181
12	0.532	0.661	**27**	0.367	0.47	**300**	0.113	0.148
13	0.514	0.641	**28**	0.361	0.463	**400**	0.098	0.128
14	0.497	0.623	**30**	0.349	0.449	**500**	0.088	0.115
15	0.482	0.606	**35**	0.325	0.418	**1000**	0.062	0.081

Таблица 16*П*

Минимальные значения коэффициента ранговой корреляции Спирмена, достоверно отличные от нуля ($df = n - 2$)

n	α		n	α		n	α	
	0.05	**0.01**		**0.05**	**0.01**		**0.05**	**0.01**
5	0.94		17	0.48	0.62	29	0.37	0.48
6	0.85		18	0.47	0.60	30	0.36	0.47
7	0.78	0.94	19	0.46	0.58	31	0.36	0.46
8	0.72	0.88	20	0.45	0.57	32	0.36	0.45
9	0.68	0.83	21	0.44	0.56	33	0.34	0.45
10	0.64	0.79	22	0.43	0.54	34	0.34	0.44
11	0.61	0.76	23	0.42	0.53	35	0.33	0.43
12	0.58	0.73	24	0.41	0.52	36	0.33	0.43
13	0.56	0.70	25	0.40	0.51	37	0.33	0.42
14	0.54	0.68	26	0.39	0.50	38	0.32	0.41
15	0.52	0.66	27	0.38	0.49	39	0.32	0.41
16	0.50	0.64	28	0.38	0.48	40	0.31	0.40

ТЕМАТИЧЕСКИЙ УКАЗАТЕЛЬ

Анализ
– дисперсионный 50
– корреляционный 61
– регрессионный 72
Баллы 10
Варианта 10
Вариационный ряд 12
Величина признака 15
Вероятность 27
– доверительная 29
– статистическая 27
Выборка 10
Выборочная оценка 15
– – средней 15
– – дисперсии 16
Генеральная совокупность 27
Гистограмма 12
Градации 52
Дисперсионный анализ 50
Дисперсия 16
– общая 51, 76
– остаточная 76
– регрессии 76
– случайная 51
– факториальная 51
Доверительная вероятность 29
– зона регрессии 74
Доверительный интервал 31
Достоверность отличий 37
Закон больших чисел 30
Изменчивость признака 11
– сопряженная 61
Исключение вариант 34
Корреляция 60
– качественных признаков 69
– ложная 63
– множественная 64
– рангов 67
– частная 65
Коэффициент вариации 19
– детерминации 77
– корреляции 61
– регрессии 72
Криволинейная регрессия 77
Критерий непараметрический 41
– параметрический 41
– χ^2 Пирсона 43
– t Стьюдента 38
– U Уилкоксона 41
– Q Розенбаума 43
– F Фишера 39, 52
Линия регрессии 72
Медиана 16
Метод наименьших квадратов 72
– φ Фишера 39
Мода 16
Нулевая гипотеза 6, 34
Объем выборки 13
Отклонение нормированное 34
– среднее квадратическое 16
– стандартное 16
Оценка параметра 27
Ошибка параметра выборки 30
– репрезентативности 30
– средней арифметической 30
– статистическая 30
Параметры распределения 15
Плотность распределения 27
Показатель корреляции рангов 68
Преобразование долей (φ) 39
– переменных 78
– z 62
Признаки дискретные 10
– качественные 10
– количественные 11
– непрерывные 11
Проба 11
Проверка гипотезы 35
Размах изменчивости 13
Ранг 41
Ранжирование 41
Распределение 12
– альтернативное 24

– биномиальное 21
– двумерное 60
– логнормальное 19
– нормальное (Гаусса) 20
– полиномиальное 25
– Пуассона 22
– равномерное 26
Регрессия 71
– линейная 71
– криволинейная 77
Репрезентативность 30
Сила влияния фактора 51
Случайная величина 4
Сравнение выборок 37
– дисперсий 39
– долей 39
– распределений 43
– средних 37
Средняя арифметическая 15
– – взвешенная 16
Статистика 35
Статистические задачи 4
Статистический вывод 8
Степени свободы 37
Уравнение регрессии 72
– аллометрическое 78
– степенное 78
– криволинейное 78
– линейное 72
Уровень значимости 29
Частость 21, 27
Частота 12
Частота теоретическая 44
– эмпирическая 44
Экспресс-метод 15, 18
Эллипс рассеяния 60

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 3
Принципы биометрии 3
Этапы биометрического исследования 4

ВЫБОРКА 10
Признак 10
Варьирование 11
Построение вариационного ряда 12

ВЫЧИСЛЕНИЕ ПАРАМЕТРОВ ВЫБОРОК 15
Средняя арифметическая 15
Стандартное отклонение 16

ОСНОВНЫЕ ТИПЫ РАСПРЕДЕЛЕНИЙ ПРИЗНАКОВ 19
Нормальное распределение 20
Биномиальное распределение 21
Распределение Пуассона 22
Альтернативное распределение 24
Полиномиальное распределение 25
Равномерное распределение 26

СТАТИСТИЧЕСКАЯ ОЦЕНКА ГЕНЕРАЛЬНЫХ ПАРАМЕТРОВ 27
Свойства нормального распределения 27
Генеральная совокупность 29
Ошибка репрезентативности выборочных параметров 30
Доверительный интервал 31
Определение точности опыта 32
Оптимальный объем выборки 33

ОЦЕНКА ПРИНАДЛЕЖНОСТИ ВАРИАНТЫ К ВЫБОРКЕ 34

ОЦЕНКА РАЗЛИЧИЙ ДВУХ ВЫБОРОК 37
Сравнение средних арифметических 37
Сравнение долей 39
Сравнение показателей изменчивости 39
Сравнение выборок с помощью непараметрических критериев . . . 41
Критерий U Уилкоксона – Манна – Уитни 41
Критерий Q Розенбаума 43
Сравнение двух частотных распределений. Критерий хи-квадрат . 43

ОЦЕНКА ВЛИЯНИЯ ФАКТОРА 50
Однофакторный дисперсионный анализ количественных признаков 51
Непараметрический однофакторный дисперсионный анализ 54
Двухфакторный дисперсионный анализ количественных признаков . 55

ОЦЕНКА ЗАВИСИМОСТИ МЕЖДУ ПРИЗНАКАМИ 59
Корреляционный анализ 61
Ложная корреляция 63
Множественная корреляция 64
Частная корреляция 65
Ранговая корреляция 67
Коэффициент контингенции 69
Регрессионный анализ 71
Линейная регрессия 71
Криволинейная регрессия 77

ВМЕСТО ПОСЛЕСЛОВИЯ 81

СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ 84

ПРИЛОЖЕНИЕ. СПРАВОЧНЫЕ ТАБЛИЦЫ 85

ТЕМАТИЧЕСКИЙ УКАЗАТЕЛЬ 101

Учебное издание

Ивантер Эрнест Викторович
Коросов Андрей Викторович

ЭЛЕМЕНТАРНАЯ БИОМЕТРИЯ

Учебное пособие

2-е издание

Редактор О. В. Обарчук
Компьютерная верстка А. В. Коросова
Оформление обложки Ю. М. Коросовой

Подписано в печать 16.02.2010. Формат 70 х 100 $^{1}/_{16}$.
Бумага офсетная. Печать офсетная.
Уч.-изд. л. 5.7.
Тираж 300 экз. Изд. № 31.

Государственное образовательное учреждение
высшего профессионального образования
ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Отпечатано в типографии Издательства ПетрГУ
185910, Петрозаводск, пр. Ленина, 33